A Bayesian network approach to poultry science: discovering connections between genetic features and stress in chickens

Emiliano Ariel Videla Rodríguez

Universitat de Lleida

TESI DOCTORAL

A Bayesian network approach to poultry science: discovering connections between genetic features and stress in chickens.

Emiliano Ariel Videla Rodríguez

Memòria presentada per optar al grau de Doctor per la Universitat de Lleida
Programa de Doctorat en Salut

Director/a
V. Anne Smith

Tutor/a
Loreta Mª Medina Hernández

2022

Acknowledgements.

I would like to thank my supervisor, Dr. V. Anne Smith, my second supervisor, Dr. John B.O. Mitchell, and my tutor, Dr. Loreta Mª Medina Hernández for standing along the journey. I am really grateful for all the opportunities and possibilities they have offered me to explore and develop my academic career.

I would like to thank the ChickenStress European Training Network for giving me the opportunity to be a member of the network. I am grateful for the chance to acquire new skill, improve the ones I already had, and build networks with other PhD students and academic and non-academic members.

I would like to thank Dr. Per Jensen, Dr. Carlos Guerrero Bosagna, and Dr. Fábio Pértille for their contribution to my thesis and their predisposition.

I would like to thank the University of St. Andrews for the opportunity to work at their facilities, as I had the chance to meet principal investigators, professors, and other PhD students.

I would like to thank the University of Lleida for giving me the opportunity to carry out my doctoral studies. I would also like to thank the external reviewers for reading the thesis and sending me comments and suggestions to improve the quality.

Finally, I would like to thank my family and my friends for the support throughout these past years.

Publications derived from the thesis.

The chapter **"MICROARRAY, BRAIN, AND STRESS"** corresponds to a manuscript submitted for publication in December 2021 as: Videla Rodriguez E.A., Mitchell, J.B.O., and Smith V.A. Application of Bayesian networks to unravel interactions between genes related to stress in chickens (Gallus gallus). *Poultry Science.*

The chapter **"RNA-SEQ, SPLEEN, AND STRESS"** corresponds to a manuscript published in Scientific Reports: Videla Rodriguez E.A., Mitchell, J.B.O., and Smith V.A. A Bayesian network structure learning approach to identify genes associated with stress in spleens of chickens. *Scientific Reports*, 12, 7482.

The chapter **"EPIGENETICS, BRAIN, AND STRESS"** corresponds to a manuscript recently accepted for publication (14-06-2022) in BMC Bioinformatics: Videla Rodriguez E.A., Pértille F., Guerrero-Bolsagna C., Mitchell, J.B.O., Jensen P., and Smith V.A. Practical application of a Bayesian network approach to poultry epigenetics and stress. *BMC Bioinformatics*.

Index

Abbreviations. ***6***

Genes. ***8***

Figure Index. ***10***

Table Index. ***13***

Abstract ***15***

Resumen ***16***

Resum ***17***

1. Introduction. ***18***

1.1. Poultry industry and animal welfare. 19
1.2. Stress response and its effects on the physiology. 22
1.3. A bioinformatic approach to stress. 27
1.4. Bayesian networks. 30
1.5. Bayesian networks and their application in poultry science. 35
1.6. Final considerations. 48

2. Objectives and Hypothesis. ***51***

2.1. Hypothesis. 52
2.2. General objective. 53
2.3. Specific Objectives. 53

3. Methods. ***54***

3.1. Data collection. 55
3.2. Data discretisation and the creation of a list of arcs to be blocked. 57
3.3. Bayesian network structure learning. 60
3.4. Approaches to the consensus Bayesian network. 65
3.5. Biological interpretation of the results. 66

4. Microarray, brain, and stress. ***68***

4.1. Introduction. 69
4.2. Dataset. 70
4.3. The Bayesian network approach. 72
4.4. Results and partial discussion. 76

5. RNA-sequencing, spleen, and stress. ***81***

5.1. Introduction. 82
5.2. Dataset. 84
5.3. The Bayesian network approach. 85
5.4. Results 90
5.5. Partial discussions 92

6. Epigenetics, brain, and stress. ***98***

6.1. Introduction. 99
6.2. Dataset. 100
6.3. The Bayesian network approach. 106
6.4. Results 110
6.5. Partial discussions 118

7. Discussion. ***122***

7.1. Chicken vs human-related studies. 123
7.2. Dealing with a challenging field of research. 125
7.3. A bigger picture of the biology behind the results. 129

7.4. Future directions in the genomic field. 134
7.5. Limitations and applications outside genetics. 136

8. Conclusions. 139

9. References. 142

Abbreviations.

ACTH: Adrenocorticotrophin Factor.

AIC: Akaike's Information Criterion.

BDe: Bayesian Dirichlet equivalent.

BIC: Bayesian Information Criterion.

BN: Bayesian Networks.

C-POD: Cancer Prevention and Outcomes Data.

CORT: Corticosterone.

CRH: Corticotrophin Releasing Factor.

DAG: Direct Acyclic Graph.

DAVID: Database for Annotation, Visualization, and Integrated Discovery.

DEG: Differentially Expressed Genes.

DMR: Differentially Methylated Region.

DNA: Deoxyribonucleic Acid.

ENA: European Nucleotide Archive.

GABA: Gamma-Aminobutyric Acid.

GEO: Gene Expression Omnibus.

HPA: Hypothalamic-Pituitary-Adrenal.

HSP: Heat Shock Proteins.

LUCADA: English Lung Cancer Database.

MDA: Malondialdehyde.

MDL: Minimum Description Length.

MI: Mutual Information.

NCBI: National Center for Biotechnology Information.

QTL: Quantitative Trait Loci.

RNA: Ribonucleic Acid.

SANS: Sympathetic-Adrenergic Nervous Systems.

SNP: Single Nucleotide Polymorphisms.

TBARS: Thiobarbituric acid reactive substances.

Genes.

ACC: Acetyl-Coenzyme A Carboxylase.

AMPK: Adenosine Monophosphate-activated Protein Kinase.

ANN: Artificial Neural Networks.

Apo-B: Apolipoprotein B.

APTX: Aprataxin.

BAG3: BAG cochaperone 3.

BRAT1: BRCA1 associated ATM activator 1.

BRCA1: BRCA1 DNA repair associated.

C20orf96: Chromosome 20 open reading frame 96.

CANX: Calnexin.

CARD19: Caspase Recruitment Domain family member 19.

CH25h: Cholesterol 25-hydroxylase.

CPT-1: Carnitine Palmitoyltransferase 1.

CRELD2: Cysteine Rich with EGF Like Domains 2.

CYGB: Cytoglobin.

DNAJA4: DnaJ Heat Shock Protein family (Hsp40) member A4.

EPN3: Epsin-3.

FAS: Fatty Acid Synthase.

FBN1: Fibrillin 1.

GHR: Growth Hormone Receptor.

H-FABP: Heart-type Fatty Acid-Binding Protein.

HSP90B1: Heat Shock Protein 90 Beta family member 1.

HSPA4L: Heat Shock Protein family A (Hsp70) member 4 like.

HSPH1: Heat Shock Protein family H (Hsp110) member 1.

IGF: Insulin-like Growth Factor.

LEP-R: Leptin Receptor.

L-FABP: Liver-type Fatty Acid-Binding Protein.

OCLN: Occludin.

PDE1C: Phosphodiesterase 1C.

PK: Pyruvate Kinase.

POMC: Proopiomelanocortin.

PSLG-1: P-Selectin Glycoprotein Ligand-1.

RNPC3: RNA Binding Region (RNP1, RRM) containing 3.

SERPINA10: Serpin family A member 10.

SPV: Support Vector Machine.

StAR: Steroidogenic Acute Regulatory protein.

THRSP-α/ THRSP-β: Thyroid Hormone responsive spot 14.

TNNT3: Troponin T3, Fast Skeletal type.

TPST: Tyrosyl Protein Sulfotransferase.

UCP: Uncoupling protein.

XPO1: Exportin 1.

Figure Index.

Figure 1. Direct acyclic graph. Graphical representation of a Bayesian network with 5 variables. Note the direction of the arrows do not represent causation, but instead a statistical relationship. (adapted from (Felipe, Silva, Valente, & Rosa, 2014)).

Figure 2. Bayesian network and the Markov Blanket property. The BN is composed by 9 variables and 8 arcs between them. The Markov Blanket property was applied to the Variable of interest (VoI; nodes belonging to the Markov Blanket of VoI are highlighted in blue). This property uses the set of parents (Variable 1 and Variable 2), children (Variable 3), and spouses (Variable 4) to make the VoI completely independent from the rest of the network. Note the direction of the arrows do not represent causation, but instead a statistical relationship. Yellow node = Variable of Interest, blue nodes = Markov Blanket of the VoI, grey nodes = nodes that do not have any possible interaction with the VoI.

Figure 3. Visual representation of two of the most common discretisation methods. Panel A shows the equal interval method, which divides the data into bins of the same range of values: in this example, data was divided into 3 bins of same length, 5 data points were assigned to the 0-2 bin, 3 data points were assigned to the 3-5 bin, and 4 data points were assigned to the 6-8 bin. Panel B shows the equal quantile method, which divides the data into bins containing the same number of observations: 4 data points were assigned to each one of the three bins (0-1, 2-5, and 6-8).

Figure 4. Graphical representations of the three types of connections in Bayesian networks. Panel A: the graph displays the convergent connection, where variable 3 has two incoming arcs. Panel B: the graph displays the serial connection, where variable 3 has one incoming and one outgoing arc. Panel C: the graph displays the divergent connection, where variable 3 has two outgoing arcs.

Figure 5. Number of genes as a function of their adjusted P-values. A total of 1397 probes with showing differential expression patterns according to the condition (control vs stress) had P-values equal to or lower than 0.05. The red line shows the threshold (threshold = 0.02) used to select the 31 probes for further analyses.

Figure 6. Visualisation of presence or absence of arcs of 50 consensus Bayesian networks learnt in Banjo. The heatmap shows the arcs found by Simulated Annealing while searching the space, visiting a total of 250 million networks. Each column represents a consensus Bayesian network built by combining the top 100 highest scoring networks. Rows represent arcs found by the algorithm. Those arcs present in at least 50% of the networks (threshold = 25 out of 50 networks) were selected to build the weighted network.

Figure 7. Steps taken and decisions made to build the weighted Bayesian network (BN). The starting point was a dataset consisting of 3 chickens under control and 3 chickens under heat stress conditions. Bioinformatic analyses were performed to normalize, to correct the background noise, and to identify differentially expressed genes (DEG). A total of 31 probes showing differential expression patterns were searched in four other datasets sharing the same animal model (chicken), the same tissue (brain), and the same high-throughput technology (microarray). Each dataset was individually discretized into three-state variables and then merged into a larger dataset consisting of 46 observations and 25 DEGs (12 probes coded for

6 genes (2 “duplicated” probes x 6 genes = 12 probes; 31 probes – 12 “duplicated” probes = 25 DEGs); the corresponding expression values of each one of these “duplicated” probes were averaged into one single value by duplicated probe). The software Banjo was utilised to learn discrete BNs, exploring the search space with a simulated annealing and the BDe score, visiting a total of 250 million networks. An initial consensus BN was built by combining the top 100 highest scoring networks. Heatmaps were used to visualize the results different consensus BNs and due to variation in the final sets of arcs, 50 consensus BN were further combined into a weighted network, by selecting those arcs present in at least 50% of the consensus BNs (threshold: 25 out of 50 networks).

Figure 8. Bayesian network corresponding to highly significant genes related to stress. Nodes correspond to genes, while edges represent the relationship between genes. The network was built considering the edges present in at least 25 out of 50 consensus networks. Note the direction of the arrows do not represent causation, but instead a statistical relationship.

Figure 9. Visualization of P-values per dataset. Number of probes (A) and genes (B) according to their P-values per dataset. Top: GSE119387, bottom: GSE85434; n represents the number of probes showing differential expression patterns (A). Shared genes by the two datasets and their corresponding P-values. Blue bars: GSE119383, grey bars: GSE85434. Genes are represented by their ENSEMBL GENE ID coded as the first three letters and the last five numbers (e.g., ENSGALG00000001573 is coded as ENS01573) (B).

Figure 10. Visualisation of presence or absence of arcs of 100 consensus Bayesian networks learnt in Banjo. The heatmap shows the arcs found by Simulated Annealing while searching the space, visiting a total of 250 million networks. Each column represents a consensus Bayesian network built by combining the top 100 highest scoring networks. Rows represent arcs found by the algorithm. Those arcs present in at least 50% of the networks (threshold = 50 out of 100 networks) were selected to build the weighted network.

Figure 11. Steps taken and decisions made to build the weighted Bayesian network (BN). The starting point was a set of two studies evaluating stress in the spleen of chickens. Bioinformatic analyses were performed to normalize, to correct the background noise, and to identify genes with differential expression patterns in each dataset. A common “stress signal” was identified between the two datasets: 19 shared genes. The augmented dataset consisted of 50 observations, 19 genes, and as both datasets evaluated stress, the stress condition was included as a binary variable (control = 0; stress = 1). The software Banjo was utilised to learn discrete BN, exploring the search space with a simulated annealing and the Bde score, visiting a total of 250 million networks. An initial consensus BN was built by combining the top 100 highest scoring networks. Heatmaps were used to visualize the results different consensus BNs and due to variation in the final sets of arcs, 100 consensus BN were further combined into a weighted network, by selecting those arcs present in at least 50% of the consensus BNs (threshold: 50 out of 100 networks). Considering that the stress was included as the variable of interest, the Markov Blanket property of BN was applied to identify key genes related to stress. Additionally, communities of densely connected nodes were identified to increase the number of genes closely related to the stress condition.

Figure 12. Bayesian network and community analysis of a set of genes. 19 genes showing differences in expression pattern were initially included at the time of learning the structure of the network in addition to the stressful condition; however, only 16 out of those 19 were linked in a network structure. Nodes represent each one of the genes and the stressful condition (circle-

shaped node, thick outline), the edges represent probabilistic dependencies between the nodes. Note the direction of the arrows do not represent causation, but instead a statistical relationship. The Markov Blanket of the stress condition (rectangle-shaped nodes) consisted of two genes, CARD19 (child) and CYGB (spouse). Five communities of densely connected nodes were identified (different colours represent different communities). The community of the condition consisted of 4 genes (CARD19, EPN3, CYGB, and BRAT1, highlighted in pink).

Figure 13. Steps taken and decisions made to build a consensus Bayesian network. The starting point was a dataset consisting of 46 chickens, 22 raised under control conditions and 24 raised under stress conditions. Bioinformatic analysis were performed as described in (Pértille et al., 2017, 2020). Thereafter, a set of 60 differentially methylated regions (DMRs) were selected based on a p-value equal to 0.005. The corresponding methylation values of each DMR were counts (values ranged between 0 and 39). A binary discretization method was implemented, considering that the most frequent value was 0. The software R (and Rstudio) was utilised to learn discrete BN. Specifically, the bnlearn package was used, exploring the search space with a score-and-search algorithm and the Bde score. A contingency test (chi-square test) was applied to all possible pairs of variables to create a list of links to avoid, considering that the data had imbalances between the binary states that could lead to the discovery of artefactual links that should not be part of the consensus network. By using the software BayesPiles, it was possible to decide that the search space was complex and building the consensus Bayesian network required a strategic and accurate approach: the combination of a model averaging and the selection of arcs common to all searches into the weighted BN.

Figure 14. Distribution of four of the differentially methylated regions (DMRs) once a binary discretization method was applied. The state 0 represents values with absence of methylation, the state 1 represents values with presence of methylation. These four DMRs are representative of imbalances between the two states, as zero was the most popular state among different DMRs.

Figure 15. BayesPiles investigation of search space. Top networks found from four separate collections of searches, representing peaks of many different hills in the search space. BayesPiles visualises a summary of network structure as a shaded stack representing out-degree of each node (darker=higher) above a bar representing network score (longer=higher), with networks along the x-axis and nodes along the y-axis. (A) shows the highest 25 networks for four collections of searches (different colours), with highest-scoring network to the left. The strong variation in network structure (different patterns in the shaded bars) indicates that these networks are tops of different peaks in the search space, not the final climb of a single hill. (B) shows the final 25 networks from all four searches combined, sorted by their score. The mixing of colours throughout shows the high variation in search peaks: each collection of searches explored different areas of the search space, finding different high-scoring structures.

Figure 16. Consensus networks of DMRs. Networks were built with common arcs to 50 searches, each one of these searches consisted of a starting point of 100 random graphs. Features representing the differentially methylated regions (named by related gene or region, see Methods) and the stress conditions are nodes; lines between nodes represent the identified relationships. Note the direction of the arrows do not represent causation, but instead a statistical relationship. Arc labels represent the average probability of belonging to the consensus network, the higher the values, the higher the probability of belonging to a high scoring network. Different colours represent different ranges of probabilities: black: 0.90-1.00, blue: 0.89-0.80; grey: 0.79-0.70; orange: 0.69-0.60.

Table Index.

Table 1 Functional Annotation Table provided by the Database for Annotation, Visualization, and Integrated Discovery (DAVID) corresponding to the Heat Shock Proteins interacting with other four genes. Terms particularly relevant to the stress condition are highlighted in bold. The gene symbol is underlined in parenthesis.

Table 2. Functional Annotation Table provided by the Database for Annotation, Visualization, and Integrated Discovery (DAVID) corresponding to the four genes found to be in close relationship with the stressful condition. Terms particularly relevant to the stress condition are highlighted in bold. The gene symbol is underlined in parenthesis.

Table 3. Differentially methylated regions and their annotations. List of differentially methylated regions (DMR) with their corresponding genetic annotation terms. The first column ("SYMBOL") represents the abbreviated gene name of the methylated region; those which say "annotated" plus a number means that the symbol for that particular DMR was not available; the second column ("Gene ID") represents the ENSEMBL gene ID; the third column ("Description") represents the description of the DMR (NA for those not available); the fourth column ("Type of DMR") represents the type of DMR (e.g., Promoter, Intron, etc.); and finally, the fifth column ("SYMBOL network") represents the name used in the consensus Bayesian network.

Table 4. Arcs and their corresponding probabilities of being part of a high scoring network. List of arcs identified between differentially methylated regions and with the stress condition, with their corresponding probabilities of being part of a high scoring network. The first column ("arc") is an arbitrary numbering for the arc; the second column ("from") represents the parent node for each arc (arcs from); the third column ("to") represents the child node for each arc (arcs to); the third column ("Average.Probablity") represents the average probability value for each arc of being part of a high scoring network.

Abstract / Resumen / Resum

Abstract

Understanding biological systems can be a very difficult and challenging task. Stress can be considered as one of these biological systems: it is a complex phenomenon which perception and integration starts in the brain, triggering the stress response with many implications on performance, health status, and welfare of poultry species. In the field of genetics and epigenetics, hundreds of thousands of genetic/epigenetic markers can be measured per individual between non-stressed and stressed birds. The genetics of stress can be further explored by the application of Bayesian network (BN) algorithms with the aim of identifying hallmark genetic features associated with stress as well as unravelling hidden interactions between them. BNs are directed acyclic graphs that represent the joint probability distributions of a given set of variables; they consist of a set of nodes, which represent the variables, and a set of arcs or edges, representing the relationships between nodes. In this context, the aims of the current thesis were to collect data from different studies and to identify a reduced number of genetic features associated with stress in chickens (*Gallus gallus*), unravelling informative relationships and interactions by the implementation of BN algorithms. Two genetic and one epigenetic datasets evaluating the effects of stress in chickens were explored. The biology behind our findings showed that genes and epigenetic variables pointed towards chaperon-related activity of Heat Shock Proteins, apoptotic and DNA damage pathways together with wound healing mechanisms, and adherent and tight junction functionality. These genes could be crucial under the exposure to stress, especially those playing a role in keeping the correct functioning of other proteins or structures. To consider these genes as key biomarkers of stress would require further research with short- and long-term goals, working towards improving the health and welfare of poultry species in terms of management and breeding programs.

Resumen

Comprender sistemas biológicos puede ser una tarea difícil y desafiante. El estrés puede ser considerado como uno de estos sistemas biológicos: es un fenómeno complejo cuya percepción e integración comienza en el cerebro, gatillando la respuesta de estrés con muchas implicancias en la productividad, el estado de salud, y el bienestar de aves de corral. En el campo de la genética y epigenética, cientos de miles de marcadores genéticos/epigenéticos pueden ser cuantificados por individuo entre aves no estresadas y estresadas. La genética del estrés puede ser explorada en profundidad mediante la aplicación de algoritmos de redes Bayesianas (BN) con el fin de identificar variables genéticas claves asociadas con el estrés como así también para descubrir interacciones ocultas entre ellas. Las BNs son gráficos acíclicos dirigidos que representan las distribuciones de probabilidades conjuntas de un conjunto definido de variables; consisten en un conjunto de nodos, que representan las variables, y un conjunto de flechas, que representan las relaciones entre nodos. En este contexto, los objetivos de esta tesis fueron recolectar datos provenientes de otros estudios e identificar un conjunto reducido de variables genéticas asociadas al estrés en pollos (*Gallus gallus*), descubriendo relaciones e interacciones informativas mediante la aplicación de algoritmos de BNs. Dos conjuntos de datos genéticos y uno epigenético que evaluaron los efectos del estrés en pollos fueron explorados. La biología detrás de los datos demostró que los genes y variables epigenéticas están relacionadas a la actividad de Proteínas de Estrés Térmico, vías apoptóticas y de daño del ADN en conjunto con mecanismos de reparación de tejidos, y la funcionalidad de uniones adherentes. Estas variables genéticas podrían resultar cruciales bajo condiciones de estrés, especialmente aquellas con un rol de protección del correcto funcionamiento de otras proteínas o estructuras. Considerar a estos genes como marcadores clave del estrés requeriría estudios futuros a corto y largo plazo, con el fin de mejorar la salud y el bienestar de aves de corral en cuestión de manejo y programas de cría.

Resum

Comprendre sistemes biològics pot ser una tasca difícil i desafiadora. L'estrès pot ser considerat com un d'aquests sistemes biològics: és un fenomen complex la percepció i la integració del qual comença al cervell, gatillant la resposta d'estrès amb moltes implicacions en la productivitat, l'estat de salut, i el benestar d'aus de corral. En el camp de la genètica i epigenètica, centenars de milers de marcadors genètics/epigenètics poden ser quantificats per individu entre aus no estressades i estressades. La genètica de l'estrès pot ser explorada en profunditat mitjançant l'aplicació d'algorismes de xarxes Bayesianes (BN) per tal d'identificar variables genètiques claus associades amb l'estrès i descobrir interaccions ocultes entre elles. Les BN són gràfics acíclics dirigits que representen les distribucions de probabilitats conjuntes d'un conjunt definit de variables; consisteixen en un conjunt de nodes, que representen les variables, i un conjunt de fletxes, que representen les relacions entre nodes. En aquest context, els objectius d'aquesta tesi van ser recol·lectar dades provinents d'altres estudis i identificar un conjunt reduït de variables genètiques associades al'estrès en pollastres (Gallus gallus), descobrint relacions i interaccions informatives mitjançant l'aplicació d'algorismes de BNs. Dos conjunts de dades genètiques i un d'epigenètic que van avaluar els efectes de l'estrès en pollastres van ser explorats. La biologia darrere de les dades va demostrar que els gens i les variables epigenètiques estan relacionades a l'activitat de Proteïnes d'Estrès Tèrmic, vies apoptòtiques i de dany de l'ADN en conjunt amb mecanismes de reparació de teixits, i la funcionalitat d'unions adherents. Aquestes variables genètiques podrien resultar crucials sota condicions d'estrès, especialment aquelles amb un rol de protecció del funcionament correcte d'altres proteïnes o estructures. Considerar aquests gens com a marcadors clau de l'estrès requeriria estudis futurs a curt i llarg termini, per tal de millorar la salut i el benestar d'ocells de corral en qüestió de maneig i programes de cria.

1. Introduction.

1.1. Poultry industry and animal welfare.

Animal production, including poultry industry, has increased during the last few decades, as a consequence of the population growth, the higher demand for animal-derived products, such as meat, milk, or eggs, as well as socio-economic interests (Moekti, 2020; Windhorst, 2006). The extensive domestication process that production animals have undergone together with the availability of better technologies oriented towards improving animal welfare have had an impact not only on production parameters but also on the quality of life of production animals (Dixon, 2020; Ericsson & Jensen, 2016; Løtvedt, Fallahshahroudi, Bektic, Altimiras, & Jensen, 2017; Moekti, 2020; Windhorst, 2006). In particular, the poultry industry breeds different species such as chickens (*Gallus gallus*), turkey (*Meleagris gallopavo*), and quail (*Coturnix coturnix*). The domestication process of these poultry species might have started 8000 years ago, in Eastern Asia (Løtvedt et al., 2017; Tixier-Boichard, Bed'Hom, & Rognon, 2011). During the beginning of this domestication process, birds adapted to live in close relationship with humans, reducing the fear response against them and, therefore, increasing tameness. These groups of less fearful and more tame birds started to reproduce in the new environment close to humans, favouring the selection of domesticated phenotypes (Bélteky, Agnvall, Johnsson, Wright, & Jensen, 2016; Price, 1999; Tixier-Boichard et al., 2011). It is not yet clear whether the appearance of the domesticated phenotypes is a consequence of the highly driven artificial selection of these traits, or it is a secondary effect of the domestication process itself (Bélteky et. Al, 2016; Price, 1999; Tixier-Boichard et al., 2011). Despite this fact, the domesticated phenotype of poultry species such as chickens or turkeys can be distinguished from its ancestor by considering some traits, such as weight, plumage colour, and some other more complex traits related to reproduction and/or behaviour (Bélteky et al., 2016; Tixier-Boichard et al., 2011). The artificial selection of poultry species for productive traits has led to the appearance of commercial breeds as they are known

nowadays, e.g., broiler chickens are reared for their meat, while laying hens chicken are reared for their eggs.

Initially chickens, in particular hens, were mostly raised in plain battery cages, where food and water were provided ad libitum but with no access to a nest where they can lay eggs or a perch where they can perform natural behaviours. However, the consumers of animal-derived products in general, started to demand for improvements of the breeding conditions where animals were raised with the aim of improving the welfare of production animal. As a consequence, plain battery cages were banned in the European Union, leading to the appearance of new housing systems such as enriched battery cages, or aviaries with or without access to an outdoor garden (Lay et al., 2011; Moekti, 2020; Philippe et al., 2020; Sosnówka-Czajka, Herbut, & Skomorucha, 2010). Birds gained access to an environment where they could perform a wider range of behaviours, due to the presence of perches, ramps, nests, tiers, and in some cases (such as free-range breeding systems), an outdoor space. In the context of animal welfare, animals must be raised under the premisses of two main paradigms, contemplating a series of aspects that need to be satisfied, such as nutrition, absence of thermal or physical discomfort, disease, injury, pain, fear, distress, the possibility to express natural behaviours, and the mental or psychological state of animals (Mellor, 2016; Webster, 2016). These two paradigms are known as the Five Freedoms or the Five Domains of animal welfare (Mellor, 2016; Webster, 2016). On the one hand, the Five Freedoms paradigm, proposed by Webster (2001) considers the following premisses:

- Freedom from thirst, hunger, and malnutrition: By providing access to fresh water and a diet appropriate to maintaining optimal health and vigour.
- Freedom from discomfort and exposure: By providing an appropriate environment including shelter and a comfortable resting area.

- Freedom from pain, injury, and disease: By prevention or rapid diagnosis and treatment of potential diseases.
- Freedom from fear and distress: By ensuring conditions and treatment which avoid mental suffering.
- Freedom to express normal behaviour: By providing sufficient space, proper facilities, and company of the animal's own kind.

On the other hand, the Five Domains paradigm, proposed by Mellor (Mellor, 2016; Mellor & Beausoleil, 2015), includes two main domains, a physical/functional domain and an affective experience domain, that can be further divided into five new domains, nutrition, environment, health, behaviour and mental state, (Mellor, 2016; Mellor & Beausoleil, 2015). In addition, each one of the five domains is divided into positive or negative aspects:

- Nutrition:
    - Negative: restricted water and food; poor food quality.
    - Positive: enough water and food; balanced and varied diet.
- Environment:
    - Negative: uncomfortable or unpleasant physical features of environment.
    - Positive: physical environment comfortable or pleasant.
- Health:
    - Negative: disease, injury, and/or functional impairment.
    - Positive: healthy, fit, and/or uninjured.
- Behaviour:
    - Negative: behavioural expression restricted.
    - Positive: able to express rewarding behaviours.

- Mental state:
    - Negative experiences: thirst, breathlessness, anger, frustration, hunger, pain, boredom, helplessness, malnutrition malaise debility, weakness, loneliness, depression, chilling, overheating, nausea, sickness, anxiety, fearfulness, hearing discomfort, dizziness, panic, exhaustion.
    - Positive: pleasure of drinking, calmness, health, fitness, reward, maternally rewarded, satiety, goal-directed, exited playfulness, physical comforts, social interactions, sexually gratified.

It is important to mention that both paradigms can be used in different but complementary approaches to assess the welfare of animals raised in productive environments (Webster, 2016). According to Webster (2016), the Five Domain approach might be more orientated to evaluate the psychological and mental state of the animals considering the physical, nutritional, and social environment, while the Five Freedoms approach can be used to assess the impacts of environmental changes that promote a better animal welfare (Webster, 2016).

1.2. Stress response and its effects on the physiology.

Independently of the paradigm of animal welfare, several aspects of animal welfare are being considered by the poultry industry when birds are raised with production aims. However, the environmental conditions where poultry species are bred might still be potentially stressful, threatening the freedom or domain related to distress or environment, respectively. Stress can be defined as a non-specific response of the organism to either external or internal stimulus, and a stressor can be defined as an external or internal factor that triggers the stress response (Lara & Rostagno, 2013; Selye, 1950; Siegel, 1971). From the moment poultry species hatch, they are exposed to a series of potentially stressful conditions, such as transportation from the

hatchery to the rearing farm where they will be raised, food regimens, social hierarchies and social interactions with con-specifics, aggressive behaviours, temperatures above or below those optimal for development, among many others (Bowling, Forder, Hughes, Weaver, & Hynd, 2018; Calefi, Quinteiro-Filho, Ferreira, & Palermo-Neto, 2017; Cantet, Yu, & Ríus, 2021; Fernandez, Labaque, Orso, Marin, & Kembro, 2021; Gasparino et al., 2013; Lay et al., 2011; Miller & Mench, 2006; Quinteiro-Filho et al., 2010; Wickramasuriya et al., 2022). The potentially stressful environmental conditions are perceived and integrated at the interplay of two main systems, the nervous and the endocrine system, also known as the neuroendocrine interplay. This interplay is composed of four main axis: the Sympathetic-Adrenergic Nervous Systems (SANS), the Hypothalamic-Pituitary-Adrenal (HPA) axis, the vagal-cholinergic axis and the Gut-Brain axis (Adelman & Martin, 2009; Ashley & Demas, 2017; Davison, 2014; Kuenzel & Jurkevich, 2010; Webster, Marketon & Glaser, 2008). Once the stressful stimulus is integrated, the stress response is triggered and involves two different responses: i) a general, non-specific response that is independent of the nature of the stressor; and ii) a specific response according to the nature of the stressor. The general, non-specific stress response involves mostly the SANS and the HPA axis, releasing a different set of stress mediators such as catecholamines and glucocorticoids, with the aim of dealing or coping with the stressful condition (Calefi et al., 2017; Selye, 1950; Siegel, 1995). Within seconds, epinephrine and norepinephrine, released by the SANS, promote the "fight or flight" response, increasing blood pressure, muscular tone, heart rate as well as the availability of metabolic resources such as glucose in the short term (Kuenzel & Jurkevich, 2010; Siegel, 1995). Within minutes, glucocorticoids, mostly corticosterone (CORT) in birds, are released as a consequence of the activation of the HPA axis. Initially, the corticotrophin releasing factor (CRH), is segregated by the hypothalamus, reaching its target gland, the anterior pituitary, stimulating the release of adrenocorticotrophin factor (ACTH). The ACTH is released into the bloodstream, reaching its

target gland, the adrenal glands, stimulating the release of CORT (Calefi et al., 2017; Scanes, 2016; Siegel, 1995). CORT promotes a relatively longer response, involving several physiological, metabolic, and behavioural adjustments with the aim of dealing with the influence of the stressor in the long term (Adelman & Martin, 2009; Calefi et al., 2017; Siegel, 1980, 1995). On the other hand, the specific stress response depends on the nature of the stressor. For example, the exposure of birds to high environmental temperatures demands an increased breathing rate and panting, as well as spreading the wings as a behavioural adaptation to dissipate excess heat (Siegel, 1971, 1995). As another example, when birds are in close proximity to a potential predator, they can either display an escaping behaviour, running as fast as possible from the predator, or they can display a tonic immobility response, also referred as freezing behaviour, staying immobile until the threat is no longer in sight (Gallup, 1977; Jones, Satterlee, & Ryder, 1992).

The stress response has been exhaustively studied in birds, evidencing the impacts on several aspects of physiology, metabolism, and behaviour (Cantet et al., 2021; Dhabhar, 2009; Elfwing et al., 2015; Fernandez et al., 2021; Hedlund, Palazon, & Jensen, 2021; Kembro, Satterlee, Schmidt, Perillo, & Marin, 2008; Quinteiro-Filho et al., 2010; Shini, Kaiser, Shini, & Bryden, 2008; Siegel, 1995; Wickramasuriya et al., 2022). The exposure of birds to different types of stressful conditions, such as thermal stress, food stress, and immune stress, among many others, has proven to alter development and growth, the plasmatic concentrations of metabolites, the innate and acquired immune responses, the gastrointestinal function, the microbiota, and many others (Burkholder, Thompson, Einstein, Applegate, & Patterson, 2008; Cantet et al., 2021; Dong et al., 2007; Fernandez et al., 2021; Lin, Sui, Jiao, Buyse, & Decuypere, 2006; Malheiros, Moraes, Collin, Decuypere, & Buyse, 2003; Mashaly et al., 2004; Mumma, Thaxton, Vizzier-Thaxton, & Dodson, 2006; Nazar, Videla, & Marin, 2018; Quinteiro-Filho et al., 2010; Shi et al., 2019; Wickramasuriya et al., 2022). Genetic and

epigenetic changes associated with a stressful condition have also been evaluated, considering the domestication process that poultry species have undergone (Elfwing et al., 2015; Fallahshahroudi, Løtvedt, Bélteky, Altimiras, & Jensen, 2019; Fallahsharoudi et al., 2017; Long, Gianola, Rosa, Weigel, & Avendaño, 2009). For example, White Leghorns chickens, a particular breed that lays white eggs, and the closest wild species to the putative chicken ancestor, the Red Junglefowls, exposed to the same physical restraint stressor for 15 minutes showed differences between the expression patterns of some genes, including StAR, CH25h, and POMC (Fallahsharoudi et al., 2015). StAR and CH25h are involved in the metabolism of Cholesterol, and therefore it is directly connected with the first steps of the stress response cascade; POMC is involved in several aspects of the adrenal functioning, including the production of hormones (Fallahsharoudi et al., 2015). Another study related to stress and domestication suggested that two genes, SERPINA10 and PDE1C, might be candidates genes associated with differences between the stress responses of the ancestral and the domesticated chicken (Fallahsharoudi et al., 2017).

In spite of the detrimental effects of stress, it is important to consider that within a group of birds there might be different responses to the same environmental conditions, where some birds might be more susceptible to the stress condition than others. Stress resilience is the capability of the organism to deal effectively, physiologically and behaviourally, with a particular stressful condition, and the capability to recover the baseline state, independently of the nature of the stressor (Pfau & Russo, 2015; Ross, Rausch, Vandenberg, & Mason, 2020). Stress resilience is a complex phenomenon, involving several brain circuits, transcriptional and epigenetic mechanisms, as well as external environmental factors and stimuli that play an important role in the resilience to stress (Pfau & Russo, 2015; Ross et al., 2020). From an individual point of view, birds (and animals in general) can be divided into two main categories based on the responses to a stressful stimulus based on their coping style. Proactive birds will

show a lower activation of the HPA axis, leading to low levels of CORT, and consequently, these birds will be more active and less fearful to take risks. On the other hand, birds with a reactive coping style will have a strong activation of the HPA, leading to high levels of CORT, and as a result, these birds will be more passive and more fearful, avoiding risks (Campbella, Hincha, Downing, & Lee, 2016; Cockrem, 2007; de Haas, Kops, Bolhuis, Groothuis, Ellen, & Rodenburg 2012; Pusch, Bentz, Becker, & Navara 2018). When it comes to coping styles and stress resilience in groups of birds, Nazar et al (2015), for example, showed that quail display two different types of responses based on the plasmatic CORT levels, where quail with high levels of CORT had higher frequency of leukocyte distribution and levels of Interleucine-13, and low percentage of inflammation against a mitogen, titres of antibodies against a non-pathogenic antigen, gamma interferon, and Interleucine-1. On the other hand, quail with low levels of CORT had low frequency of leukocytes and levels of Interleucine-13, but high percentage of inflammation against a mitogen, titres of antibodies against a non-pathogenic antigen, gamma interferon, and Interleucine-1 (Nazar, Barrios, Kaiser, Marin, & Correa, 2015). Differential susceptibilities to stress have also been shown between breeds. For example, the exposure to heat stress of two genetically different breeds (White Leghorns and Fayoumi) used for egg production showed that the stressor differentially affected physiological blood parameters, such as sodium, carbon dioxide partial pressure, bicarbonate, glucose, among others (Wang et al., 2018). Additionally, the differences were seen not only between breeds but also within birds (independently from the breed), and according to the authors, the Fayoumi birds showed stronger and more elastic physiological responses to heat stress (Wang et al., 2018). It is important to mention that Fayoumi birds have undergone a selection programme for heat tolerance, being more resistant to heat stress, while White Leghorns are more susceptible to stress due to the artificial selection (Wang et al., 2018). Another study considering three different breeds, Red Junglefowl, Village fowl, and broiler breeders (reared

for their meat), also showed different susceptibilities to heat stress, the broiler breeders being more susceptible to heat stress as evidenced by an increase in body temperature, the heterophil to lymphocyte ratio, and the corticosterone concentrations (Soleimani, Zulkifli, Omar, & Raha, 2011).

1.3. A bioinformatic approach to stress.

The stress response as well as stress resilience can be evaluated on several aspects of the physiology of the organism, genetics being one of them. The recent advances in technology have allowed the possibility to measure genetic features, such as gene expression, or epigenetic modifications with microarrays or RNA-sequencing, as well as the accurate identification of changes in the patterns of genetic features according to the environmental conditions by the development of powerful statistical tools. Microarrays and RNA-sequencing are also known as high-throughput technologies because they measure hundreds of thousands of genes within a biological sample. Both types of technologies work with cDNA transcripts, obtained from the initial RNA samples, but with different approaches. Microarray technologies combine hundreds of thousands of DNA probes that have been attached to a glass surface. Then, the target sequences hybridize with the probes, and by the fluorescence of a signal bound to the target sequences plus the implementation of scanning and processing software, the expression of thousands of individual genes is measured (Mantione et al., 2014; Rao et al., 2019). RNA-sequencing, on the other hand, needs the RNA to be initially fragmented into short sequences before obtaining the cDNA transcripts. Then, an adapter is added at the end of the fragments before amplifying the sequences. These steps that transform the RNA into cDNA are known as library preparation. Thereafter, the cDNA segments are sequenced by labelled nucleotides, and finally, the resulting sequences are counted (Mantione et al., 2014; Rao et al., 2019). Even though these two technologies have different approaches, they both are used to measure gene

expression Epigenetic modifications of the DNA sequence can also be measured as indicators of the activation of the stress response. However, changes in this field of research, epigenetics, are not in terms of the DNA sequence itself but considering external modifications of the DNA. Epigenetic modifications of the DNA sequence include the addition of molecules to nucleotides sequences, such as methyl groups, or the way the DNA is structured and compacted by histones, that might have an impact on the levels of expression a gene can have (Pértille et al., 2017, 2020; Skinner, Manikkam, & Guerrero-Bosagna, 2010).

Despite the genetic or epigenetic features, in general, working with high-throughput technologies involves working with hundreds of thousands genes (Greene, Tan, Ung, Moore, & Cheng, 2014; Shendure & Ji, 2008). These types of technologies allow the evaluation of a large set of genetic features with only one sample per individual, usually comparing the differences between one or more conditions such as stress (control vs stress), sex (male vs female), domesticated phenotypes (non-domestic vs domestic), among many others (Elfwing et al., 2015; Ericsson et al., 2016; Fallahsharoudi et al., 2017). However, dealing with hundreds of thousands of genes, as variables or features of interest, might be challenging, not only because of the amount of data collected but also of the way data should be analysed (Greene et al., 2014). As a consequence, the appearance of biological big data demanded new ways of analysing complex biological systems and informatic approaches (termed 'bioinformatics'), applied statistics, and machine learning techniques started to be implemented on biological data (Fallahshahroudi et al., 2019; Greene et al., 2014; Løtvedt et al., 2017; Pértille et al., 2017; Tarca, Carey, Chen, Romero, & Drăghici, 2007; Zahoor, De Koning & Hocking, 2017). In general, the starting point consists of removing background noise and normalising the raw data originated by the high-throughput technologies (Bélteky et al., 2016; Greene et al., 2014; Guo et al., 2020; Pértille et al., 2020; Xia et al., 2019; Zilliox & Irizarry, 2007). The next step is to identify highly significant genetic or epigenetic features by the application of statistical tools,

such as t-tests, that compare the values from two conditions, such as control vs stress, and determine differences in the expression or epigenetic patterns (Goerlich, Nätt, Elfwing, Macdonald, & Jensen, 2012; Pértille et al., 2017; Saelao et al., 2018). After these series of steps, the outcome is a list of genes than can be further annotated with the aim of exploring their biological functionality with tools such as KEGG pathways, GO terms, or other databases such as those provided by the National Center for Biotechnology Information. It is then possible to understand whether the genes are involved in metabolic pathways, transduction cascades, immune responses, or several other physiological processes. However, one of the main drawbacks of this approach is that the list of highly significant genes or epigenetic features could be quite extensive, and they could be related to numerous aspects of the physiology of the organism, such as immunity, cell signalling, neurogenesis, among many others. Fallahsharoudi et al. (2015) found a total of 1291 transcripts that were differentially expressed between the domestic and the ancestral chickens. Some of the genes were related to receptors, such as GABA, channels, as well as steroidogenesis. Guo et al. (2020) found around 300 differentially expressed genes in the spleen of chicken exposed either to control or to stress conditions. These genes were mostly related to the responses of the immune system, involving processes associated with biological processes, cellular components, and molecular function. Kuchipudi et al (2014) exposed chicken and duck cells to an immune challenge, the influenza virus, identifying thousands of differentially expressed genes. These genes were divided depending on whether they were related to general biological processes or specifically related to the immune system. Genes were related to pathways such as signal transduction, enzymatic processes, transcription activities, immune signalling (Kuchipudi et al., 2014). In terms of epigenetic features, differences in the methylation patterns of the DNA sequence have been identified between two conditions, such as control vs social isolation stress, or two poultry housing systems, such as cages and aviaries (Pértille et al., 2017, 2020). The identified

epigenetic features were related to processes related to cell signalling pathways, the immune system, cardiac conduction, among others (Pértille et al., 2017, 2020).

In this context where an extensive list of genes or epigenetic features are proved to have different patterns depending on a particular condition of interest, a further step can be taken with the aim of mining the biological data. Mathematical and computational tools can be implemented to classify, predict, or unravel patterns within biological big data (Greene et al., 2014; Tarca et al., 2007). The tools are also known as machine learning algorithms, and they can be divided into two main categories, supervised or unsupervised, with the main difference between them being the type of data available (Greene et al., 2014; Tarca et al., 2007). In order to work with supervised algorithms, the data must have at least one variable with labels or classes (e.g., control vs stress, male vs female, domestic vs ancestral), and these techniques are used to build models which aim to either classify or predict the behaviour of a new set of observations but with unknown classes. Examples of supervised algorithms are linear regression, linear discriminant analysis, support vector machines (SPV), artificial neural networks (ANN), Bayesian networks (BN), among others (Greene et al., 2014; Tarca et al., 2007). The unsupervised algorithms do not require the data to have a variable with labels or classes associated to the individuals; these algorithms are used to unravel or discover hidden patters within the data, as well as to identify groups of variables or individuals that can be part of a cluster. Examples of unsupervised algorithms are correspondence analysis, principal component analysis, and clustering analysis (Greene et al., 2014; Tarca et al., 2007). This particular thesis will be especially focused on the application of BNs, as an exploratory method to discover relationships and interactions within a set of genetic and epigenetic features considering chickens reared either under a control or a stressful condition.

1.4. Bayesian networks.

BNs have been defined as graphical models that allow a concise representation of the probabilistic dependencies between a given set of random variables as a directed acyclic graph (DAG). In other words, BNs display the joint probability distribution of a given set of variables (Heckerman, Geiger, & Chickering, 1995; Nagarajan, Scutari, & Lèbre, 2013a, 2013b; Needham, Bradford, Bulpitt, & Westhead, 2007). The DAG consists of a set of nodes, representing each one of the variables measured to a specific experimental unit (chickens in the case of this thesis), and a set of arcs, edges, or links, representing the connections between the nodes (Heckerman et al., 1995; Nagarajan et al., 2013a, 2013b; Needham et al., 2007). Considering the definition of BNs, the structure of the network is acyclic because the mathematical properties of a joint probability distribution cannot loop back on itself (Heckerman et al., 1995; Nagarajan et al., 2013a, 2013b; Needham et al., 2007). Figure 1 represents a simplified graphical representation of a BN, with five variables. This simple BN can also be seen as a family tree, where, for example, a particular node of interest can have parents, children, and/or spouses. Bearing this in mind, the following information can be extracted from the network, focusing on the "variable of interest" (Felipe et al., 2014; Nagarajan et al., 2013a, 2013b; Needham et al., 2007):

- The variable of interest (VoI) has two parents, Variable 1 and Variable 2, evidenced by the incoming arcs from Variable 1 and Variable 2 to the VoI.
- The VoI has one child, Variable 3, evidenced by the outgoing arc from the VoI to Variable 3. It is important to mention that the VoI can also have two or more children.
- The VoI has one spouse, Variable 4, evidenced by the common child (Variable 3) shared between the VoI and Variable 4.

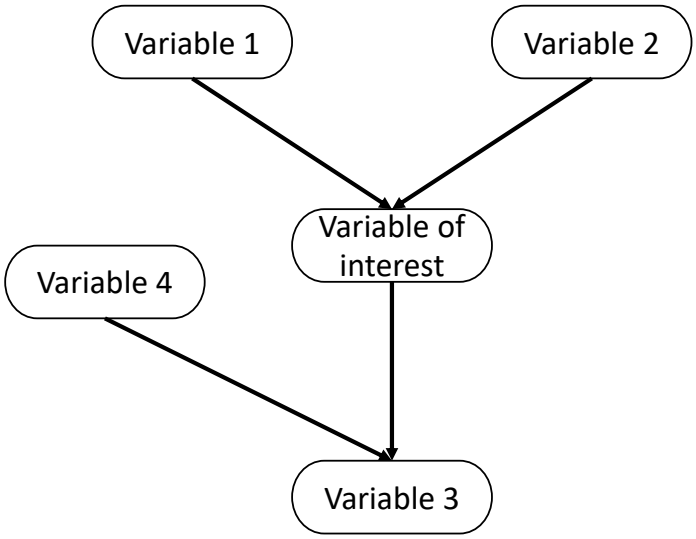

Figure 1. Direct acyclic graph. Graphical representation of a Bayesian network with 5 variables. Note the direction of the arrows do not represent causation, but instead a statistical relationship. (adapted from (Felipe et al., 2014)).

BNs are based on probability theory, therefore, considering a set of variables, $(X_1, X_2, \dots, X_n)$, and the previously described DAG, the following formula can used to describe the network (Heckerman et al., 1995; Pearl, 1988):

$$\Pr(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} \Pr\,(Xi|Pa_i)$$

where *n* represents the number of variables, Pa_i represents the set of parents of Xi in the direct acyclic graph (Felipe et al., 2014; Nagarajan et al., 2013a, 2013b). A variable without parent/s will have a probability of its value given the null set (e.g. $\Pr(X_i)$). Focusing on Figure 1, the previous formula could be applied as follows (Pr stands for probability; V1, V2, V3, V4, and VoI represents Variable 1, Variable 2, Variable 3, Variable 4, and Variable of Interest respectively):

Pr(V1, V2, VoI, V3, V4) = Pr(V1) Pr(V2) Pr(VoI| V1,V2) Pr(V3 | VoI, V4) Pr(V4)

Considering the formula as well as Figure 1, V1, V2, and V4 are conditionally independent, while, on the other hand, VoI is conditionally dependent given the values of V1 and V2, and V3 is conditionally dependent given the values of VoI and V4 (Felipe et al., 2014). Considering that BNs can be seen as a family analogy, one of their main properties is known as the Markov Blanket, and it allows the possibility to put the focus on a variable of interest (Figure 2) (Aliferis, Statnikov, Tsamardinos, Mani, & Koutsoukos, 2010; Felipe et al., 2014; Heckerman et al., 1995; Nagarajan et al., 2013a, 2013b; Needham et al., 2007). This property uses the set of parents, children, and spouses to make the VoI completely independent from the rest of the variables that do not belong to the Markov Blanket (Aliferis et al., 2010). Figure 2 shows a BN composed by 9 variables (including the VoI) and 8 arcs between the given set of variables. Applying the Markov Blanket property to the VoI allows the possibility to reduce the number of nodes related to a particular variable (the VoI in this case): 4 out of 8 variables are closely associated to the VoI, while Variables 5, 6, 7, and 8 (grey nodes) do not have any possible interaction with VoI. In this thesis, working with the Markov Blanket property represents the possibility to further reduce the number of genetic features associated with the stress condition.

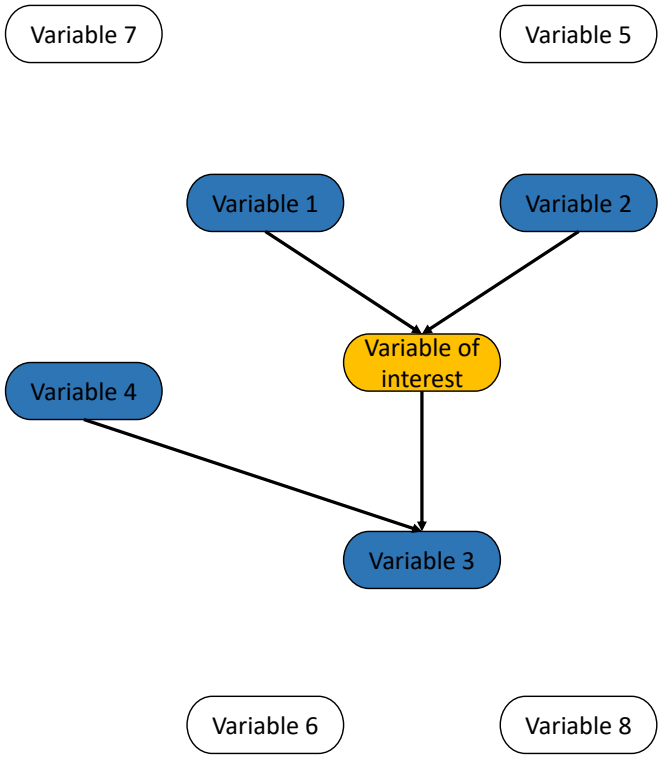

Figure 2. Bayesian network and the Markov Blanket property. The BN is composed by 9 variables and 8 arcs between them. The Markov Blanket property was applied to the Variable of interest (VoI; nodes belonging to the Markov Blanket of VoI are highlighted in blue). This property uses the set of parents (Variable 1 and Variable 2), children (Variable 3), and spouses (Variable 4) to make the VoI completely independent from the rest of the network. Note the direction of the arrows do not represent causation, but instead a statistical relationship. Yellow node = Variable of Interest, blue nodes = Markov Blanket of the VoI, grey nodes = nodes that do not have any possible interaction with the VoI.

In terms of the applications of BNs, they can be used with three different goals: i) structure learning, ii) parameter inference, and iii) variable inference. The first application of BNs is related to learning the structure of the network itself. The structure of the network can be learnt using previous knowledge and expert opinions, it could be learnt from the data themselves, or a combination of both approaches (Heckerman et al., 1995; Milns, Beale, & Smith, 2010; Needham et al., 2007; Parsons et al., 2005). This application of BNs allows the possibility to extract crucial information from the overall structure of the network, initially identifying relationships and interactions between variables (Balov, 2013; Hidano et al., 2015; Li, Wu, Zhang, & Yang, 2010). In the field of genetics and epigenetics, relationships between variables can be useful to reveal or understand gene pathways, as they could bring light into

new associations between genes related to particular metabolic processes (Balov, 2013; Li et al., 2010). Another useful application of structure learning is the implementation of the Markov Blanket property as an approach to identify highly connected nodes or to focus on a particular node of interest, potentially considering these highly connected nodes as relevant for a particular condition (Li et al., 2010; Xia et al., 2019). The next application is parameter inference, that refers to the process of learning the conditional probabilities of a BN given the data and/or prior information. Bearing in mind the formula previously described, the probability of each node is calculated based only on the probabilities of its parent nodes. When all the parameters (probabilities) are learnt, Conditional Probability Tables can be built for local nodes and for the global structure of the BN (Heckerman et al., 1995; Kyrimi et al., 2021). This application is also referred to as parametrization of the network, and once it is done, the last application can be used: variable inference. This particular application is based on a particular piece of evidence, and it allows the possibility to query the BN according to the given evidence. For example, and as a hypothetical case, the structure of the BN involves variables related to heart attacks, such as hypertension, smoking, obesity, physical exercise, among others, and a particular piece of evidence corresponding to a patient indicates that they suffer from hypertension. Providing the network with this piece of evidence (hypertension = yes), the rest of the BN will display the conditional probabilities for each node, with special emphasis on heart attack, potentially increasing the probabilities of suffering a heart attack (Kyrimi et al., 2021; Nagarajan et al., 2013a, 2013b).

1.5. Bayesian networks and their application in poultry science.

BNs analyses have been applied in different aspects of poultry industry. Previous studies can be divided into two main categories, whether the aim of the study was focused on the individual (e.g., chickens or quail), or whether the focus was on the poultry industry itself

(e.g., overall structure of the poultry industry or the food chain management). This section will try to give an overall view of how BNs have been implemented in different aspects related to poultry science.

Considering the three types of applications of BNs, most the studies focused on learning the structure of the BNs. Researchers were looking to get insights into the relationships and interactions between a wide variety of variables. Variables involved were related to egg production (Felipe et al., 2014), gene expression (Hidano et al., 2015; Li et al., 2010), antimicrobial resistance (Hartnack et al., 2019), housing systems and management (Comin, Jeremiasson, Kratzer, & Keeling, 2019; Hartnack et al., 2019). The other application of BNs in poultry studies was to build an expert system, learning the structure of a BN with a given set of variables and data points, and then using it with new entries (combining structure learning, parameter inference, and variable inference). These expert systems were built to make decisions such as classifying the egg freshness (Soltani & Omid, 2015), predicting the apparent metabolizable energy of the food (Alvarenga et al., 2021), or predicting the condemnation risk of a flock (Lupo et al., 2013). Among the studies, there were those evaluating the performance of BNs and comparing it to other machine learning tools utilised, based on some performance parameters (Felipe et al., 2014; Long et al., 2009; Parsons et al., 2005; Soltani & Omid, 2015). The BNs were compared to some commonly used models, such as linear regression, SPV, ANN, Naïve Bayes, and decision trees (Felipe et al., 2014; Long et al., 2009; Soltani & Omid, 2015). Additionally, one study compared BNs with other two complex models implementing Markov chain Monte Carlo and simulations (Parsons et al., 2005). Regarding the aims of the studies, they were varied in nature: some studies evaluated the performance of BNs as machine leaning tools, testing BNs as models to predict the outcome of new entries and compare its performance with other methods (Felipe et al., 2014; Long et al., 2009; Parsons et al., 2005; Soltani & Omid, 2015). Other studies had the aim of getting insights into a particular field of

research, mostly focusing on the structure of the BN and the unravelled informative interactions and relationships between the variables (Comin et al., 2019; Hartnack et al., 2019; Hidano et al., 2015; Li et al., 2010).

Before learning the structure of the BNs, some of the studies required pre-processing of the data using discretization methods, dimensionality reduction, and the creating of lists of arcs to be blocked. For example, Hidano et al. (2015) when studying antimicrobial resistance, the phenotypes associated to antimicrobial resistance were coded as a binary variable (high-resistance or low-resistance) (Hidano et al., 2015). Long et al. (2009), when evaluating the performance of BNs, Naïve Bayes, and ANN, discretized the class, mortality of chicks, into 2, 3, 4, 5, and 10 categories based on a k-means algorithm. Similarly, Lupo et al. (2013) discretized some continuous variables into two-state categories (low or high), while others were discretized following either a decision tree or a k-means algorithm (Long et al., 2009; Lupo et al., 2013). Due to the dimensionality of the initial datasets, some of the studies utilised feature selection algorithms to select highly important variables to the class (Long et al., 2009; Soltani & Omid, 2015). These algorithms implemented either a correlation-based (Soltani & Omid, 2015) or an information gain-based feature selection algorithm (Long et al., 2009). When working with correlation, researchers were looking for variables highly correlated with the variable of interest (but low correlation values between variables); while the information gain algorithm was based on the difference in entropy before and after removing a particular variable associated with the class (Long et al., 2009; Soltani & Omid, 2015). Some researchers in their studies created a list of arcs to be blocked based on prior knowledge before learning the structure of the BNs. The aims were i) to block arcs between variables that did not have a biological meaning or could not be possible from a biological point of view, and ii) to simplify the network (Faverial, Cornet, Paul, & Sierra, 2016; Hidano et al., 2015). Finally, in terms of the software used to learn the BNs, some studies used programming languages such as

MATLAB or R (with R packages “bnlearn” and “abn” commonly used) (Alvarenga et al., 2021; Felipe et al., 2014; Hartnack et al., 2019; Hidano et al., 2015; Long et al., 2009), while others implemented software such as WEKA or Norsys Software of Vancouver (Parsons et al., 2005).

The following two section will describe more in depth the previously mentioned studies.

1.5.1 Individual animal level studies.

To start with those studies focused on individuals, Felipe and colleagues (2014) were interested in comparing the efficiency of different machine learning techniques, such as BNs, a regression model, and artificial neural networks, to predict the total egg production of European quail (Felipe et al., 2014). A set of variables, available on two different lines of female quail, combining productive and egg quality traits were used as a dataset. Productive variables included body weight per week from birth to 35 days of age, weight gain from birth to 21 days of age and from 21 to 35 days of age, age at first egg and number of eggs produced from 35 to 80 days of age. The variables related to the egg quality, measured in four different age points, included egg weight, yolk weight, egg white weight, eggshell weight, and egg specific gravity (Felipe et al., 2014). Additionally, they included the total egg production, considering the number of eggs produced from 35 to 260 days of age as the variable of interest for the model. The authors built two independent networks for each one of the lines using 31 variables in total. Even though the overall structure of the networks corresponding to each line were different, total egg production was, on the one hand, only conditionally dependent on the number of eggs produced from 35 days old to 80 days old, and on the other hand, it was independent from the rest of variables considered for the model (Felipe et al., 2014). The authors were able to discover the variable that can be used to predict the total egg production

of European quail together with the fact that the other productive variables are not of importance for predicting the number of eggs a quail will be producing. According to the authors, BNs could be a useful tool to establish associations between variables in a more comprehensive approach (Felipe et al., 2014).

Two studies were carried out in a field similar to the one being explored in this thesis: genetics. Long and colleagues (2009) studied SNPs and their relationships with mortality (Long et al., 2009). In this study, BNs, together with Naïve Bayes and Artificial Neural Networks, were utilised to select a set of SNPs associated with mortality. Initially, mortality (a continuous variable) was categorized into a multiclass variable by the application of a k-means algorithm, setting the numbers of categories or classes to 2, 3, 4, 5, and 10. Thereafter, the selection of the SNPs was done by a two-step approach, initially filtering a total of 5000 SNPs to a set of 50 informative SNPs considering the multiclass mortality variable. This set of 50 informative SNPs were further used to identify SNPs in close relationship with mortality by applying BNs, Naïve Bayes, and Artificial Neural Networks algorithms. Naïve Bayes can be considered as part of BNs, as it is also based on probability theory. The main difference between these two techniques is that Naïve Bayes assumes that all variables are independent from each other, and even though this assumption is not always true in real data, the algorithm is often used with good performance (Domingos & Pazzani, 1996; Kelemen, Zhou, Lawhead, & Liang, 2003; Long et al., 2009). Among the three algorithms, Naïve Bayes, with 2 or 3 mortality classes had the best performance in identifying SNPs in close relationship with mortality. While the study was focused on comparing the overall performance of different machine learning algorithms to select a set of relevant SNPs, the identified SNPs could be further used in genomic studies, especially those focused on quantitative traits associated to a single modification of the DNA sequence.

The other study related to the poultry genetics was aimed at unravelling relationships and interactions between genes previously known to be part of fatty acid metabolism, in addition to identifying the main genes with relevant effects on the metabolic path (Li et al., 2010). Li and colleagues (2010) measured the expression levels of a set of genes associated with the metabolism of fatty acids. The selected genes were divided into 4 major categories considering the role in the metabolism of fatty acids (Li et al., 2010):

- Genes related to fatty acid anabolism: Thyroid hormone responsive spot 14 (THRSP) α and β,fatty acid synthase (FAS), glutathione S-transferases, pyruvate kinase (PK).
- Genes involved in fatty acid catabolism: adenosine monophosphate-activated protein kinase (AMPK) subunits 2 and 3, acetyl-coenzyme A carboxylase (ACC), leptin, leptin receptor (LEP-R), lipoprotein lipase (LPL), and carnitine palmitoyltransferase 1 (CPT-1).
- Genes related to fatty acid transportation: liver-type fatty acid-binding protein (L-FABP), heart-type fatty acid-binding protein (H-FABP), uncoupling protein (UCP), apolipoprotein B (Apo-B).
- Genes related to fatty acid growth axis: growth hormone (GH), growth hormone receptor (GHR), and insulin-like growth factor (IGF) 1 and 2.

Once the structure of the BN was learnt, the authors focused on the overall structure of the network, identifying two genes at the top of the cascade, THRSP α and β, as well as a major node of interest, H-FABP, being connected to seven other genes (two parents, PK and L-FABP; five children, UCP, AMPKγ2, LPL, ACC, and AMPKγ3). The approach implemented by the authors allowed them to focus on the paths or the topological ordering of the nodes, considering the direction of the arcs, as well as the identification of a relevant gene in the metabolism of

fatty acids (Nagarajan et al., 2013a, 2013b). The authors gained insights into the interaction among the selected genes as well as a global overview of the metabolism of fatty acids (Li et al., 2010).

As a bridge between the two types of studies (individual-oriented *vs* poultry industry-oriented), Soltani and Omid (2015) focused their study on eggs, with the aim of building a useful system to inspect the freshness of eggs at different storage periods. The authors were looking for a system that combines the following characteristics: cheap, simple, rapid, non-destructive, and non-harmful (Soltani & Omid, 2015). As a first step, a set of variables were selected, identifying those that were contributing the most to the class while removing the irrelevant ones. A correlation-based feature selection algorithm was applied in order to select those variables that were highly correlated with the class, but with low correlation between the variables. The final set of variables consisted of dielectric measurements as a way to assess the quality of eggs, some traits related to egg dimensionality (e.g., volume and diameters) and 24 variables related to the air cell height (Soltani & Omid, 2015). The dataset was then divided into a training set and a test set, and several machine learning algorithms were tested, including BNs. Independently of the used searching algorithm (6 in total), BNs achieved a 100% accuracy in predicting the class. Together with two other machine learning methods, artificial neural networks and support vector machines, BNs had the highest performance. The authors concluded that BNs, as well as artificial neural networks and support vector machines, can be used as decision-making systems to predict egg freshness in a non-destructive approach (Soltani & Omid, 2015).

1.5.2 Industrial level studies.

In terms of BNs and their application in poultry-related studies, findings have been reported in a broad spectrum of fields. Among the studied, four main topics can be identified:

i) potentially harmful microorganisms and their association with risk assessments and antimicrobial resistance (Hartnack et al., 2019; Hidano et al., 2015; Parsons et al., 2005); ii) nutrition and the food chain processes related to meat products (Alvarenga et al., 2021; Lupo et al., 2013); iii) management of poultry industry (Comin et al., 2019); and iv) poultry green wastes and their impact on soil (Faverial et al., 2016).

To begin with, it is important to mention that harmful microorganisms are a source of concern in the poultry industry as they represent a threat to animal and human health as well as to the world-wide economy. A study carried out by Parsons and colleagues (2005) was focused on quantitative risk assessment, looking for probabilistic approaches to deal with uncertainty in order to make decisions. Even though the authors were interested in implementing three different methods, BNs being one of them, their approach to BNs is worth mentioning. Initially, they built a network considering prior knowledge provided by experts in the field of poultry industry and microbiology. Thereafter, the quantification of the parameters corresponding to each node was not feasible due to the lack of data sets with the required information. However, the authors combined data to individually learn the parameters for each node; data were collected from publicly available data sets, industry surveys, unpublished sources as well as expert opinions. Once the network was learnt, it was used as an expert system to infer the output of according to different pieces of evidence. For example, the network revealed that a lack of hygiene in multiple steps would lead to contamination and cross-contamination, with a higher probability of transporting contaminated animals. The contamination of the surface of eggs can be controlled by the disinfection with formaldehyde (Parsons et al., 2005).

Another main issue faced by the poultry industry with harmful impacts on the health of animals and human beings is antimicrobial resistance (Hartnack et al., 2019; Hidano et al., 2015). In this field, BNs approaches have been applied in both farms and poultry meat, so as

to gain further insights into antimicrobial resistance from two different but complementary perspectives. Hartnack and colleagues (2019) focused on a larger scale, as they worked with poultry farms and two different categories of variables, risk factors and antimicrobial resistance. Risk factors included variables such as the gender of the manager, the presence/absence of pets, the size of the farm, the housing system, whether the egg trays were re-used, the vaccinator in charge of the vaccination protocols, and the way dead birds were discarded. The antimicrobial resistance was measured in faeces against 7 antibiotics. Hidano and colleagues (2015), on the other hand, focused on domestic poultry products, such as meat and offal samples, and their antimicrobial susceptibility to antibiotics. Once defined the antimicrobial phenotypes, the expression of particular antimicrobial genes was measured. Even though the relationships and interactions between variables were complex, further insights into risk factors, antibiotic phenotypes, and antibiotic resistance genes were discovered and unravelled. Focusing on farms, the probabilities of ampicillin resistance were higher when the vaccines were not administered by a private service (Hartnack et al., 2019). When talking about antimicrobial resistance genes in poultry products, the BNs revealed interactions between the genes associated to antimicrobial resistance. This fact led the authors to consider the increase of antimicrobial resistance as systematic, by combining the effects of multiple genes given the discovered informative interactions and relationships, instead of at random (Hidano et al., 2015). Even though both studies highlighted the exploratory nature of the research, authors mentioned the importance of further research in fields such as bacteriology, ecology, and epidemiology, as a key aspect to have a better understanding of antimicrobial resistance (Hidano et al., 2015).

Following up with the next poultry science topic, even though the studies are not strictly related, they are contributing to understanding the food chain process of poultry meat, from the moment birds are fed until they are slaughtered. Nutrition of poultry species is one of the main

factors influencing the productive performance of broilers (Luna et al., 2019). The quality of the feed provided to broilers is crucial to properly gain weight, and it mostly depends on the apparent metabolizable energy associated with the availability of nitrogen (Alvarenga et al., 2011; Alvarenga et al., 2021). Alvarenga and colleagues (2021) mention that there are two possible ways to evaluate the energy content of the feedstuffs, either by chemical and biological testing (as well as the composition of nutritional tables when available), or by the application of statistical tools to predict the energy content depending on the chemical composition of the feedstuffs (Alvarenga et al., 2021). As an alternative to other methods, the authors applied BNs with the aim of predicting the values of metabolizable energy. To learn the structure of the network they included the content of crude proteins, ether extract, ash, crude fiber, the source of energy and protein (e.g., corn or soybean, respectively), the poultry species, and the values of apparent metabolizable energy. The overall structure of the network unravelled that all variables were in close relationship with the apparent metabolizable energy, leading the authors to the conclusion that the model is appropriate to calculate the energy content of feedstuffs. As a next step in the poultry food chain, before and after chickens are slaughtered, they go along an inspection in the search of abnormalities, dirtiness, or unhealthy birds, and the health status of the whole flock. All the information collected is also known as Food Chain Information. Lupo and colleagues (2012) were looking for an expert system that would allow them to classify the flocks into risk categories, based on the Food Chain Information. According to the authors, developing a decision-making tool would help the inspectors to adapt the sanitary inspection to a particular type of flocks (Lupo et al., 2013). The variables were divided into technical characteristics, such as the genetic strain, the density of the flock, the average weight of birds, among others; and into diseases and losses characteristics, such as digestive, respiratory or locomotory disorders, mortality, stress, among others. A variable extracted from previous records in relation to the condemnation proportion of previous flocks

was also included when learning the structure of the network (Lupo et al., 2013). In regards to the insights, the condemnation risk class was in close relationships with all the variables included in the network. However, the authors identified that three main variables contributed the most to condemnation of the flocks: mortality during the last seven days of rearing, health disorders during the last week of rearing, and the average weight at slaughter. Even though the authors highlighted that the tool did not have an overall good performance, it could be initially utilised to have a better interpretation of the whole process of Food Chain Information (Lupo et al., 2013).

Following with poultry industry, Comin and colleagues (2019) applied a BN approach so as to have a broad overview of the poultry production practices, management, and welfare status of poultry industry, focusing on some of the aspects of poultry industry, such as the housing system, the facilities, the management, and the welfare of commercial laying hens (Comin et al., 2019). The variables included in the analysis were: outer and inner biosecurity; the condition of the room where the packaging takes place; the lighting conditions and the quality of the air where animals are bred; the management of water; the quality and condition of the furnishing and litter; whether the barn was in compliance with the law and had an alarm system; whether the owner had a logbook with a summary of the number of eggs collected, the number of culled birds and the vaccination routine; the condition of feathers; and the occurrence of mites. The authors also included other variables such as whether the flock belonged to organic production, the size (number of birds) and age (in weeks) of the flock, the average monthly mortality, and the type of housing, whether the flock was raised in enriched cages, single or multi-tier systems (Comin et al., 2019). It is important to mention that a single-tier system only has one level above the ground floor, while the multi-tier system has up to five different levels above the ground, where birds can find nests, food, drinkers, perches to rest

during the night, among others. The authors came to the following conclusion after understanding and interpreting the network (Comin et al., 2019):

- Outer biosecurity and keeping records of collected eggs, culled birds and vaccination protocols would probably be classified as unsatisfactory for organic production. Organic production was associated with smaller flock sizes as well as unsatisfactory furnishing and litter conditions.
- Enriched cages were associated with a high presence of external parasites, better water management and air quality, and negatively associated with the mortality of the flock.
- Single-tier housing systems were associated with a smaller flock size as well as with a satisfactory lighting condition, when compared with multi-tier systems.

The authors gain valuable information from the network, not only for managerial and welfare aspects, but also for the consequences of the welfare program as well as the specific requirements for each particular housing system. According to the authors, better air quality associated with enriched cages might be related to the absence of deep litter and bacteria contributing to the production of ammonia. The reduced number of dead individuals in enriched cages might be related to the lower pressure of bacteria as well as parasites in this type of housing system. Age was only associated with feather condition, which relates to the fact that as birds get older, the feathers become worn. Even though some connections might be hard to unravel, such as the ones between environmental-based welfare conditions (lighting and air quality) and the managerial and housing variables, the authors suggested that improving the lighting system and taking extra care of the management of litter would have a positive impact on the environmental-based welfare indicators. The results obtained from this particular study can be used to understand the relationships among several aspects of poultry industries,

considering not only the health and welfare implications but also the insights into management and infrastructure of breeding laying hens with commercial goals (Comin et al., 2019).

One last BN approach to poultry industry was the one reported by Faverial and colleagues (2016). This particular study is focused on the manures and green wastes of some industries, poultry industry being one of them. The aim of the study was to identify factors and processes affecting the quality of composts, specially under tropical weather conditions. Faverial and colleagues divided their variables of interest into 3 main categories: 1) the type of manure, whether it came from cattle, goat, horse, or poultry litter; 2) the value of co-composting with a bulking agent, whether the manure includes green wastes or not; and 3) whether earthworms are included or not during the stabilization phase (phase that takes place after the thermophilic phase). Additionally, explanatory variables were included, such as the length of the thermophilic phase, the stabilization method, the pH, the content of total carbon, nitrogen, lignin, phosphorus, and potassium (measured at the initial blend, at the end of the thermophilic phase and at the end of the stabilization phase), and the mass losses of the same nutrients (carbon, nitrogen, lignin, phosphorus and potassium) measured at the end of the thermophilic and the stabilization phase (Faverial et al., 2016). A total of four BNs were built, divided into two groups: the first group of 2 BNs were related to organic material, including total carbon, lignin, and nitrogen content as variables. The second group of 2 BNs were related to the nutrient content, including nitrogen, phosphorus, and potassium content as variables. These two groups of 2 BNs were further divided into two individual networks, one for the thermophilic phase and another for the stabilization phase. As for the insights, the authors could extract crucial information: for both phases (thermophilic and stabilization) the final properties depended on the initial status and the effects of mass losses on the concentrations; the losses of contents were associated with each other, except for nitrogen, which content was not actually related to the loss of nitrogen itself but it was affected by the initial content of total carbon (during the

thermophilic phase); vermicomposting did not affect the losses of the final levels of the analysed properties. The length of the thermophilic phase was linked to the initial total carbon and the pH; co-composting was positively related to the length of the thermophilic phase (via the initial total carbon and the pH); co-composting also positively affected the mass, total carbon, and the losses of potassium via the pH, but negatively affected the initial phosphorus content. Even though the interrelationships might be complex, Faverial and colleagues could identify that the concentration effects and the quality of the raw material were the principal factors affecting the organic material and the nutrient content in addition to the stability of the final compost. Additionally, co-composting of manures and green wastes attenuated the effects of tropical climate on the losses of organic material by enhancing the content concentrations (Faverial et al., 2016).

The previous subsections gave an overall view of the current state of the applications of BNs in poultry science. The studies are varied in nature, using different types of variables to learn the structure of BNs, having different aims and scopes, and showing different findings. Researchers gained insights into their corresponding field within poultry science. Of particular interest for this thesis, the study developed by Li et al. (2010) is in a similar field of research as the one presented here, using a structure learning approach to BNs to learn informative relationship and interaction between a given set of genetic features.

1.6. Final considerations.

As previously described, BN approaches have been applied to gain further insights into several and varied aspects of poultry science. Researchers explored a variety of contexts, with different aims, and, in terms of the data, different types of variables. The applications involved both types of machine learning techniques: i) supervised, to classify or predict the outcome of a model, and ii) unsupervised, to understand complex relationships and interactions between a

given set of variables. Although BNs have many advantages and implications, they also have limitations and drawbacks. Some of the studies mentioned in the previous section highlighted that the quality and the quantity of the data have been identified as a one major issue when working with BNs (Hartnack et al., 2019; Hidano et al., 2015; Lupo et al., 2013; Parsons et al., 2005). In this regard, using relatively small number of data points will find some links but unfortunately will not find all possible links between the given set of variables. Even though this is one of the limitations of BNs, those links that are found while searching the space are actually true accurate links (Yu, Smith, Wang, Hartemink, & Jarvis, 2004). Researchers also raised some concerns about dealing with missing data as well as the size of the datasets, especially when learning the structure of the BNs. (Hartnack et al., 2019; Lupo et al., 2013). Additionally, the discretization of the data into categories was also mentioned as one of the drawbacks of BNs, and according to the authors, it is one important factor to bear in mind when applying this method (Comin et al., 2019; Hartnack et al., 2019; Hidano et al., 2015; Parsons et al., 2005).

In regard to this particular thesis, BNs will be applied in the field of genetics and epigenetics to unravel potential relationships and interactions between genetic features and a stressful condition in a poultry animal model: chickens. Identifying hallmark genetic features, such as genes, or particular DNA regions with epigenetic modifications, in close association with a stressful condition will contribute to a better understanding of the stress phenomenon, highlighting genetic/epigenetic elements that might be involved in stress resilience or stress resistance. It is important to mention that the nature of this thesis is mostly exploratory; however, the knowledge could be the used as the starting point of future poultry science research. The knowledge discovered throughout this thesis could have several implications in the short-term as well as in the long-term working towards improving the welfare and the health of chickens under commercial conditions. The identified genes can be used as genetic

indicators of stress, including a new perspective to the already existing set of nervous, hormonal, immunological, digestive, behavioural indicators, evaluating a bigger picture of the stress phenomenon. In the long term, genes can be part of artificial selection and breeding programs, potentially working towards stress resistance or stress resilience breeds of chickens that would be physiological better prepared to deal with challenging events related to the poultry breeding condition, and consequently, improving the welfare of chickens under those conditions.

2. Objectives and Hypothesis.

2.1. Hypothesis.

The exposure of chickens to stress has an impact on several systems within the biology of the organism. Different high-throughput technologies quantify genetic/epigenetic features between non-stressed and stressed chickens, identifying hundreds of thousands of these features. Dealing with large amount of data demands accurate tools to reduce the dimensionality, a required step prior learning the structure of BNs.

Bioinformatic techniques allow the possibility of identifying a set of genetic features with different patterns between non-stressed and stressed chickens. Thereafter, the search space is ready to implement BN algorithms that can unravel hidden relationships and interactions between genetic/epigenetic features and a stressful condition. Furthermore, the implementation of the Markov Blanket property of BNs can identify a small set of genetic/epigenetic features closely related to the stress condition.

The starting point of this journey is a list of hundreds of thousands of genetic/epigenetic features and the idea of narrowing it down to a smaller set makes it difficult to hypothesise and to predict the possible outcomes when it comes to the biology behind the findings. Due to the complexity of the stress phenomenon, the genetic features can be related to different biological systems, pathways, and/or processes in different cell types, tissues, and/or organs. Learning the structure of BNs will identify small groups of genetic/epigenetic features for further investigation.

2.2. General objective.

Further exploration of the genetics and epigenetics behind the stress response, and consequently, the stress phenomenon in a poultry animal model such as the chicken can be crucial from different perspectives, considering the biology, the health, and consequently, the productivity behind it. This thesis has the overall objective of discovering and unravelling informative relationships and interactions among genetic features and a stressful condition in a poultry science context, by the application of BN approaches. The thesis will contribute to a better understanding of poultry genetics and its relationship with the stress phenomenon, disentangling hidden patterns associated with stress resilience and stress resistance in chickens. Additionally, the thesis will contribute to the body of biological knowledge, especially related to computational biology.

2.3. Specific Objectives.

The specific objectives are:

i) To outline strategies to collect, reuse, and combine publicly available genetic datasets from online repositories, such as Gene Expression Omnibus or ArrayExpress, into larger datasets. The aim is to build datasets sharing the same animal model, the same tissue, and the same high-throughput technology to get further insights into the stress phenomenon by complementing previously published studies.

ii) To identify stress signals, in the form of differentially expressed genes or differentially methylated regions, across the previously collected datasets, by using bioinformatic tools.

iii) To unravel hidden relationships and interactions between the genetic/epigenetic features (stress signals) and a stressful condition (when possible) by building probabilistic networks using a BN approach.

3. Methods.

In this following chapter, a general overview of the methods implemented throughout the experimental chapters will be depicted. Initially, the way the data were collected will be described, followed by the pre-processing techniques and the combination of smaller datasets into larger datasets. Thereafter, the BN algorithms will be explained, concluding the chapter with the search for the biological interpretation of the results.

3.1. Data collection.

Data were collected from publicly available online repositories such as ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) or Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/). The main goals of these repositories are, on the one hand, for researchers to upload their high-throughput technology experiments and make them available to the scientific community, and on the other hand, for other researchers to download and reuse them with the aim of gaining a better understanding of a particular field of interest. In this thesis, we used key words such as microarray, RNA-sequencing, stress, chicken, *Gallus gallus* to explore the availability of resources. Before combining data into larger datasets, the experimental designs as well as the tissues evaluated were taken into consideration. The outcomes of these searches were divided into two different databases that were explored in the first two experimental chapters of this thesis: the first dataset involved brain, stress, and microarrays, while the second dataset involved spleen, stress, and RNA-sequencing. Considering these two datasets, the overall aim was to focus on the stress response, and consequently, the stress phenomenon as it is one of the main problems that poultry industry is facing nowadays. When it comes to the genetic field, poultry animal models are, at a certain point, difficult to work with, especially considering the number of individuals that are being tested for differences in the gene expression patterns. Previous studies evaluated a small sample

of individuals together with multiple factors, such as stress and domestication or stress and an immune challenge (Ericsson et al., 2016; Guo et al., 2020; Pértille et al., 2020). In terms of data, these two facts (small number of samples and multiple experimental factors) further reduce the amount of data available, bearing in mind all the possible combinations of levels belonging to different factors. Comparing stress with cancer, for example, there are more studies and more data available to collect and combine into larger datasets, but unfortunately, it is not the case for chickens. Even though BN algorithms can be applied in datasets with few data points, the aim of building larger datasets was to augment the number of observations as long as they share the same animal model (the chicken), the same tissue, and the same high-throughput technology. This will allow the possibility of increasing the robustness of the outcomes as well as identifying strong signals across studies, complementing the findings of previous studies. The dataset explored in the third experimental chapter was provided by one of the academic partners of the ChickenStress European Training Network, the Linköping University, Sweden. The data can be accessed through from the European Nucleotide Archive (ENA, www.ebi.ac.uk), under the accession number PRJEB34868. The dataset involved brain, stress, and epigenetic changes measured as the addition of methyl groups to particular regions of the genome.

Regarding the software, R and Rstudio were used to perform the pre-processing of the files, the bioinformatic analysis, and the visualization of the results. R packages forming part of Bioconductor – The Open-Source Software For Bioinformatics – were used with the aim of importing, pre-processing, and extracting the differentially expressed genes or the expression values of the corresponding genes. R packages related to network visualisation, such as "bnlearn", "Rgraphviz", "igraph", or "ggraph" were used with the aim of plotting the outcome of the BN algorithms (see section 3.3 below). The details and specificities of the arguments used for each case will be mentioned in the corresponding experimental chapters. R and

RStudio were also used to import and pre-process the files corresponding to the BN algorithms outcomes. R packages such as “tidyverse” and “stringr” were used to pre-process the files.

3.2. Data discretisation and the creation of a list of arcs to be blocked.

Once the data are in place and before learning the structure of the BNs, variables are usually discretised into a reduced number of discrete states. Discretising the data into a reduced number of states can be considered as an appropriate strategy when dealing with complexity, uncertainty, and non-linear relationships (Milns et al., 2010; Mitchell, Wallace, Smith, Wiesenthal, & Brierley, 2021; Nojavan, Qian, & Stow, 2017; Yu, Smith, Wang, Hartemink, & Jarvis, 2004). Different types of discretisation methods as well as different number of discrete states have been explored. The two most common types of discretisation methods are interval (or equal interval) and quantile (or equal quantile) (Figure 3). The main difference between them is the way data are distributed into the new states: while equal interval divides the data into bins of the same range of values (Figure 3, panel A), equal quantile divides the data into bins containing the same number of observations (Figure 3, panel B) (Nojavan et al., 2017).

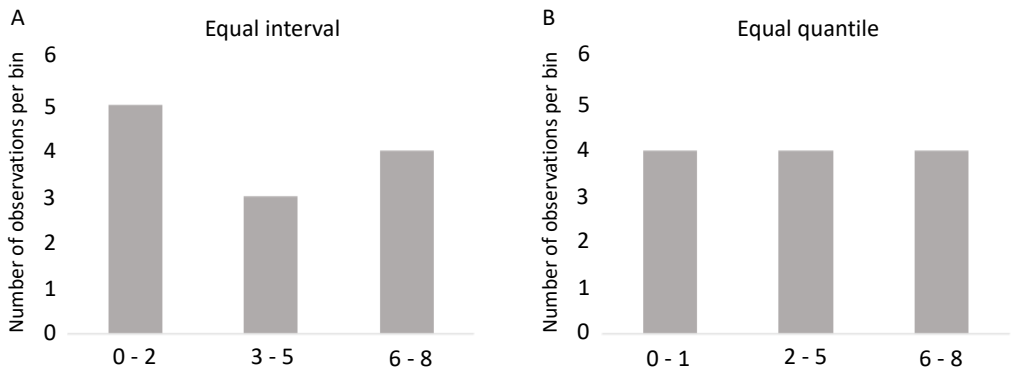

Figure 3. Visual representation of two of the most common discretisation methods. Panel A shows the equal interval method, which divides the data into bins of the same range of values: in this example, data was divided into 3 bins of same length, 5 data points were assigned to the 0-2 bin, 3 data points were assigned to the 3-5 bin, and 4 data points were assigned to the 6-8 bin. Panel B shows the equal quantile method, which divides the data into bins containing the same number of observations: 4 data points were assigned to each one of the three bins (0-1, 2-5, and 6-8).

Alternative methods have also been developed, and some other methods combine expert knowledge with a discretisation method (Chen, Wheeler, & Kochenderfer, 2017; Mitchell et al., 2021). For example, when dealing with ecological data, Mitchell et al. (2021) used an alternative discretisation method combining expert knowledge and a discretisation method based on the median. Among the variables, there were different species of zooplankton taking values that ranged from 0 to over 100. Initially, the data were divided into two categories, where values equal to zero implied the absence of the species, while any other value represented the presence of the species. This strategy could be considered as a binary discretisation method, where data are divided into two possible categories depending on their values. The next step was to further divide the data assigned to the presence category into two new categories: low counts and high counts depending on the value of the median (Mitchell et al., 2021).

Despite the discretisation method implemented, discretising continuous data into discrete states has advantages and disadvantages. One of the main drawbacks is losing information as a consequence of reducing the range of values to a relatively small number of states present in the data. However, the advantages of reducing the number of states lies on the possibility of increasing the statistical power, reducing noise, and making the process computationally less expensive (in terms of time needed to perform the analysis) (Heckerman et al., 1995; Milns et al., 2010, Apendix A and B; Yu et al., 2004). It is important to highlight that, when possible and if the data and/or previous knowledge allow it, it would be preferable to implement a discretisation method that divides the data into evenly distributed categories, such as the one implemented by equal quantiles. Previous studies suggest that statistical power would be increased at the point of finding the probabilistic relationships between a given set of variables when data are equally distributed among the categories, in addition to the previously mentioned reduction of noise (Heckerman et al., 1995; Milns et al., 2010, Apendix

A and B; Yu et al., 2004). Additionally, dividing the data points using equal quantiles is a better way to deal with the presence of outliers, considering that they will be included in either of one of the two extremes (for example, “low” or “high”). On the other hand, using an interval discretization method might consider an outlier as one of the extreme categories (for example, “low” or “high” with one single observation in it, the outlier), while the remaining values will be distributed between the other categories. However, and as previously mentioned in this section and in the introduction, discretizing the data is one of the main drawbacks of working with BNs (Comin et al., 2019; Hartnack et al., 2019; Hidano et al., 2015; Parsons et al., 2005). Bearing this in mind, in this thesis, the first choice was to implement a quantile discretisation method with three states, however, this method was not feasible to apply when working with the epigenetic dataset due to the distribution of the data. Therefore, a binary discretisation method was utilised. The details regarding the discretisation methods will be described in each one of the experimental chapters.

One of the main properties of BNs is that they allow the possibility to include prior information in the form of either a list of arcs required to be present or a list of arcs to be blocked when learning the structure of the network. The former consists of a set of arcs that should be present in the overall structure of the network, and the information could come from previous studies and/or from expert knowledge. On the contrary, the latter consists of a set of arcs that must not be included in the overall structure of the network. In this particular case, the information could either come from previous studies, expert knowledge, and/or from the implementation of a statistical tests that can identify the lack of dependencies between two given variables (Lupo et al., 2013; Milns et al., 2010; Mitchell et al., 2021). When a discretisation method is used and data are not evenly distributed into the discrete states, the creation and inclusion of a list of arcs to blocked when learning the structure of the BN must be considered, because the imbalances between the states can represent a challenge for the

algorithms, resulting in the discovery of relationships that should not be there as a result of an artefact of the algorithm itself (Milns et al., 2010, Apendix A). A previously implemented statistical test to create this list for discrete variables is a contingency test based on a chi-square test (Milns et al., 2010; Mitchell et al., 2021). The contingency test consisted on the application of a chi-square test to all possible pair of variables in order to obtain a measurable criteria to select those arcs to be banned from the network (Milns et al., 2010; Mitchell et al., 2021). A p-value equal to or higher than 0.25 was used as a threshold to create the blacklist: the pair of variables showed no possible dependence between them (Milns et al., 2010; Mitchell et al., 2021).

3.3. Bayesian network structure learning.

As mentioned in the introduction section, BNs can have three different types of applications: structure learning, parameter inference, and prediction based on a particular piece of evidence. In this particular thesis, among the three different applications of BNs previously mentioned, the main focus is on structure learning as a first approximation to the field of poultry genetics. Briefly, throughout the introduction section it was noted that a BN is a directed acyclic graph (DAG); it is acyclic because no back loops are allowed between two given variables (Heckerman et al., 1995; Pearl, 1988). It consists of a set of nodes, which represent each one of the variables, and arcs or edges establishing the relationships and interactions between them (Heckerman et al., 1995; Nagarajan et al., 2013a, 2013b; Pearl, 1988). The DAG is the representation of the joint probability distribution of a given set of variables; in other words, the probabilistic dependencies between the variables (Heckerman et al., 1995; Nagarajan et al., 2013a, 2013b; Pearl, 1988). It is possible to add some pieces of information in regards to the structure of BNs and the arcs between nodes: if we focus on three random variables, Variable 1, Variable 2, and Variable 3, there are three possible but fundamental types of connections (Heckerman et al., 1995; Nagarajan et al., 2013a, 2013b):

- Convergent connection: Variable 3 has incoming arcs from Variable 1 and Variable 2; Variable 3 conditionally depends on the joint distribution of Variable 1 and Variable 2 (Figure 4, Panel A).
- Serial connection: node Variable 3 has an incoming arc from Variable 1 and outgoing arcs to Variable 2, therefore the probability of Variable 3 is conditionally dependent given Variable 1, and the probability of Variable 2 is conditionally dependent given Variable 3, while Variable 1 is conditionally independent (Figure 4, Panel B).
- Divergent connection: node Variable 3 has outgoing arcs to Variable 1 and Variable 2, therefore the probabilities of Variable 1 and Variable 2 are conditionally dependent given Variable 3, while Variable 3 is conditionally independent (Figure 4, Panel C).

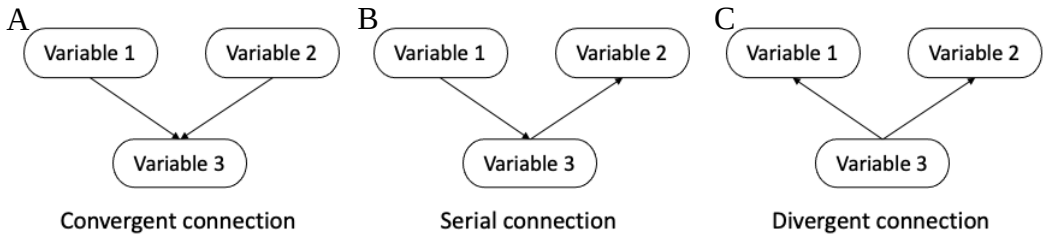

Figure 4. Graphical representations of the three types of connections in Bayesian networks. Panel A: the graph displays the convergent connection, where variable 3 has two incoming arcs. Panel B: the graph displays the serial connection, where variable 3 has one incoming and one outgoing arc. Panel C: the graph displays the divergent connection, where variable 3 has two outgoing arcs.

In terms of BN algorithms, their overall aim when learning the structure of the network is to identify the network that best fits the data. The algorithms implemented by BNs can be divided into two major categories: constraint-based or score-based (Nagarajan et al., 2013a, 2013b). In this particular thesis, the focus was on the latter ones, the score-based algorithms. These algorithms are also known as search-and-score algorithms, as they use heuristic searches to explore the space defined by the given set of variables with the aim of identifying the highest

scoring network (Heckerman et al., 1995; Nagarajan et al., 2013a, 2013b; Needham et al., 2007). The starting point is usually a random graph that the algorithm uses to initiate the exploration of the search space. The score of this first network is calculated and it is then compared with the following network the algorithm will encounter. The algorithm will try to improve the score by adding, removing, or reversing an arc in the search of the highest scoring network. Every new network is scored, and the score is then compared to the previous best scoring network, with two possible outcomes: i) if the score of the new network is lower than the previous one, the new network is discarded, and ii) if the score of the new network is better than the previous network, the new network is kept. Thereafter, the process is iterated until no further improvement on the score is made and a global maximum is reached, representing the network with the highest score (Heckerman et al., 1995; Nagarajan et al., 2013a, 2013b). Some stopping criteria could also be used such as time spent on exploring the search space or number of networks visited.

When comparing two different networks at the point of learning the structure of the network that best fits the data, BNs are based on the mathematics behind Bayes' theorem, and the probability of a graph given the data is possible to be calculated as follows:

$$\Pr(G \mid D) = \frac{\Pr(D \mid G)\ \Pr(G)}{\Pr(D)}$$

where Pr (G | D) is the probability of a graph given the data, Pr (D | G) is the probability of the data having been produced the graph, Pr (G) is the prior probability of the graph structure, and Pr(D) is the probability of the data. Comparing two networks to choose between the one that best fits the data, would mean:

$$\frac{\Pr(G_1 \mid D)}{\Pr(G_2 \mid D)} = \frac{\Pr(D \mid G_1)\ \Pr(G_1)}{\Pr(D)} \Big/ \frac{\Pr(D \mid G_2)\ \Pr(G_2)}{\Pr(D)}$$

where G_1 or G_2 represent a particular graph.

In order to calculate this, some assumptions can be made in order to compare the two graphs as follows:

I. Both networks are learnt from the same data; therefore, the term Pr (D) is the same for both equations, and they can be simplified from the equation.

II. All graph structures have equal probabilities; the prior probability term represented by Pr (G) is the same for both equations, and therefore they can be simplified from the equation.

The only term that can be calculated then is the probability of the data having been produced by the particular graph or Pr (D | G). As previously mentioned, the algorithms implemented in this thesis are known as search-and-score algorithms as they use a particular score to measure the improvements of the BNs. For example, some scoring metrics are the Bayesian Information Criterion or BIC, the Akaike's Information Criterion or AIC, the Bayesian Dirichlet equivalent score or BDe, the Minimum Description Length or MDL, among others (Lam & Bacchus, 1994; Matthäus, Smith, & Gebicke-Haerter, 2010; Yu et al., 2004). The differences between scores are based on the way they estimate the Pr (D | G), and therefore the way they define the network that best fits the data. In this particular thesis, we learnt the structure of BNs using the BDe score. The decision to use this score was made based on the work of Yu and collaborators (2004). In their study, two datasets with different number of observations (2000 data points vs 100 data points) were compared in terms of the arcs that the algorithm discovered based on two scores: the BIC score and the BDe score. When working with a small number of data points, the BDe score was able to identify some of the arcs (not the totality of them), while the BIC score did not find any of them (Yu et al., 2004). One of the limitations that the authors found was that the presence of false positives increased with low amounts of data points. However, these arcs were mostly found between nodes that did not

have a direct family relationship (e.g., parent nodes), meaning that the number of uninformative links increases. The authors discussed that the BDe score might be less conservative when dealing with a small number of data points, and this would be of importance for this thesis, as the number of observations involved in each dataset is relatively small.

To learn the structure of BNs, two open-source software applications were used: Banjo, which is available for free for academic purposes (http://www.cs.duke.edu/~amink/software/banjo/), and the R package "bnlearn". Banjo can be accessed via the terminal, and some simple command lines can be used to run the algorithms. The dataset is required to be a tab-delimited file, with rows representing the individuals and columns representing variables. On the other hand, the "bnlearn" package can be accessed by the software application RStudio. In general, BN algorithms apply heuristic searches with two possible algorithms: Greedy or Simulated Annealing, exploring the search space from empty or random graphs, scoring each network with different scores depending on whether the data are discrete or continuous. The Greedy algorithm only focuses on those arcs that increase the score, therefore, if an arc does not improve the score, it is fully discarded. On the other hand, and the main difference between the two algorithms, Simulated Annealing involves an extra parameter besides the score. Although the algorithm is searching for changes that increase the score, this extra parameter, known as temperature, represents the probability of accepting a particular change in the network that negatively affects the score, in other words, decreases the score. The higher the temperature, the higher the probability of accepting a particular arc (Matthäus et al, 2010). However, as the algorithm explores the search space, the value of the temperature tends to decrease, and as a consequence, the probability of accepting a lower scoring network decreases (de Campos & Huete, 2000; Liu & Bo, 2011).

3.4. Approaches to the consensus Bayesian network.

Considering that BN algorithms are found using heuristic searches, that add, remove, or reverse arcs modifying the structure of the current network in the search of the best scoring network, in addition to the fact that the complexity of the search space can be influenced by the type and size of the data available, it is possible that different searches can have slightly different sets of arcs (Vogogias et al., 2018). Depending on the outcome of the search, different strategies can be used in order to represent the consensus BN (Vogogias et al., 2018). In some cases, it could be possible that the algorithm is always climbing the same hill of the search space, and then, the top scoring network can be used to represent the BN that best fits the data. In other cases, it could be possible that the algorithm is not finding the same top scoring network, and then in this case, the consensus BN can be represented as the *n* best top scoring networks, *n* being a positive number such as 10 or 100. Other search spaces can be more complex and therefore they might require creative strategies to overcome this problem with the aim of representing the consensus BN. One of the approach can be to combine different consensus networks (either considered as the top scoring network or the *n* top scoring networks) into weighted networks (McNally, Heeren, & Robinaugh, 2017; Vogogias et al., 2018). Building weighted networks would mean working with a set of networks, count the number of times an arc was found across networks, and define a threshold to select those arcs forming part of the final consensus networks (McNally, Heeren, & Robinaugh, 2017; Vogogias et al., 2018). As another approach to complex search spaces, Milns et al (2010) implemented an average method taking into consideration the presence or absence of arcs coming from different searches and their scores. The model averaging implemented to deal with the complex search space compares the similarity of different networks and their scores to finally select the arcs belonging to the consensus BN (Milns et al., 2010).

In this thesis, different approaches and strategies were used in order to deal with the complexity of the search space corresponding to each one of the datasets. The details and specificities of each one of the approaches will be mentioned in the corresponding experimental chapters.

3.5. Biological interpretation of the results.

In order to provide further insights into the biological meaning of the set of identified genes, two publicly available resources were explored: the database for annotation, visualization, and integrated discovery (DAVID) (Huang, Sherman, & Lempicki, 2009), and the National Center for Biotechnology Information (NCBI). DAVID is a bioinformatic resource, publicly available, that combines different sources of information, such as protein-protein interactions, bio-pathways, GO terms, homology, literature, among many others, with the aim of closing the gap between a list of statistically significant genes and their functionality together with the biological meaning (Huang et al., 2009). A list of up to a few thousand (e.g. 2000) genetic attributes corresponding to the genes of interest is used as the starting point. DAVID will then implement some algorithms to classify the genes into groups of genes that have similar annotation terms, determine which of these genes have overrepresented biological terms, and identify related annotations and terms to a particular gene (Huang et al., 2009). DAVID provides a useful bioinformatic resource to further explore and visualize a list of genes, focusing not only on individual genes, but also on groups of genes that might be functionally related to each other based on their annotations. The outputs of DAVID are three Functional Annotation Tools; in this thesis we used the Functional Annotation Chart and the Functional Annotation Table. The former provides a list of overrepresented annotation terms considering the list of genes as a whole, including the corresponding p-value and an adjusted Benjamini p-value. The latter is focused on each individual gene, and provides their corresponding

annotation terms, which come from different sources of information, such as KEGG pathways or GO terms.

The NCBI's mission is "…to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease …coordinating efforts to gather biotechnology information…" (https://www.ncbi.nlm.nih.gov/home/about/mission/). When looking for specific genes, it might provide (depending on how much research has been carried out) information such as the Gene Symbol and annotation, the Gene ID, the location (chromosome and specific location), nucleotide sequence (FASTA files), tissue where the gene is expressed, literature, possible interactions with other genes, among many other pieces of information.

The biology behind the data corresponding to both genetic datasets (brain, stress, and microarray and spleen, stress, and RNA-sequencing) were explored initially with the DAVID bioinformatic tool, and then with the NCBI resource. The genetic attributes of each one of the genetic features were uploaded to the DAVID tool and the Functional Annotation Chart and Table provided by this tool were used to understand the biological functionality of the genetic features. Thereafter, the NCBI database was explored with the aim of getting further insights into the functionality of the genes, especially previously published studies that would guide the discussion of the findings. In regard to the epigenetic dataset, the genetic features of the topmost important DMRs were looked for in the NCBI database.

4. Microarray, brain, and stress.

4.1. Introduction.

The first approximation to the genetics of the stress phenomenon in this thesis is explored in the brain of chickens. As a starting point, the brain is where the stress response is perceived, integrated, and then triggered (De Kloet, 2003; McEwen et al., 1997; Siegel, 1971, 1995). Additionally, the central nervous system is the main point that controls, regulates, and modulates several processes such as memory, behaviour, processing, integration, and it plays a crucial role in coping or dealing with the influence of stressors (Ashley & Demas, 2017; Calefi et al., 2017; Giayetto et al., 2020; Kuenzel & Jurkevich, 2010; Nazar, Videla, Fernandez, Labaque, & Marin, 2018). Therefore, exploring the genetics of the brain under stress conditions might bring new knowledge, with future implications on the health and welfare of poultry species.

In general, genetic studies in poultry species involved a relatively small number of individuals, considering that bioinformatic analysis can be implemented in a small set of samples with the aim of identifying differentially expressed genes between non-stressed and stressed birds. Although analysing the expression patterns of genes in each one of the experiments provides with these independent sets of differentially expressed genes, there exists a wealth of information that can be extracted by combining experiments that have also collected gene expression from the same organism and tissue. This data can augment that collected in a particular designed experiment, providing further information on how genes interact. In this context, the aims of this first approximation to the stress phenomenon were to use such an augmented dataset to identify a reduced number of genes associated with a stress condition in the brain of chickens (*Gallus gallus*) that interact with each other, unravelling their relationships and interactions by the implementation of a BN approach.

4.2. Dataset.

Initially, differentially expressed genes were identified in a publicly available gene expression dataset, consisting of data from six samples, three of them coming from chickens reared under control conditions, while the other three were exposed to heat stress (GEO accession number GSE23592). According to the experimental design, chickens were assigned either to a control treatment (temperature 28±1°C) or to a heat stress treatment (temperature 40±1°C, for 3 hours), and gene expression was measured by microarray technology on brain samples. The R package "affy" (Gautier, Cope, Bolstad, & Irizarry, 2004) was used to pre-process the files, to normalise, and to correct the background noise. The R package "limma" (Ritchie et al., 2015) was used to extract the expression values, as well as to fit a linear model to identify and then select probes corresponding to a set of differentially expressed genes (DEG). A total of 1397 probes were initially identified as differentially expressed by applying the function *topTable*, considering the adjusted P-value (false discovery rate) provided by the same function to be less than or equal to 0.05. However, considering that the number of DEGs was relatively high to implement BNs, an adjusted P-value less than or equal to 0.02 was used as a threshold to select the topmost highly significant DEGs (Figure 5). This cut-off constrained the number of significant probes to 31 for further analysis.

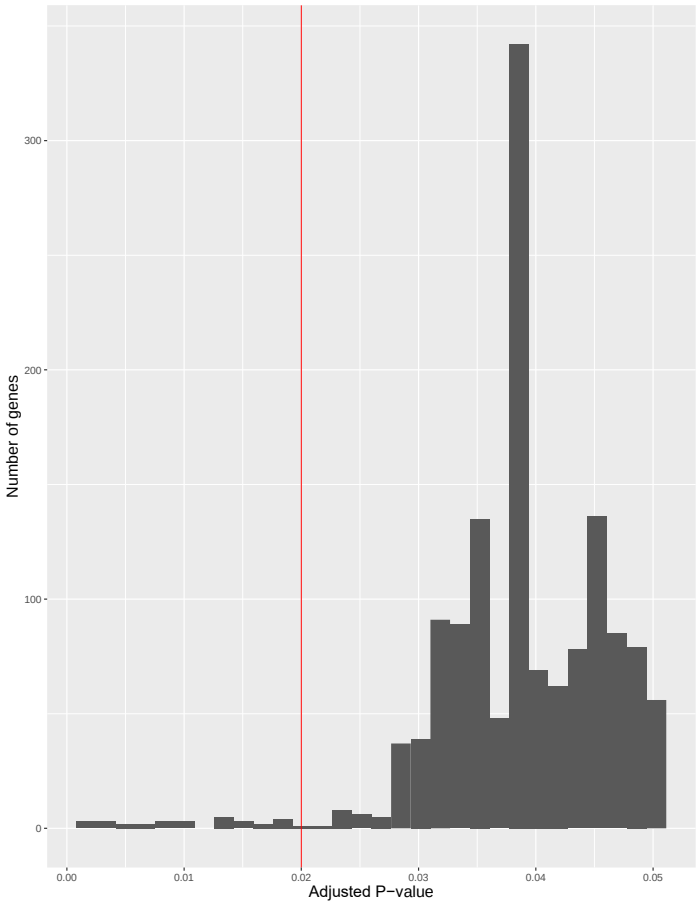

Figure 5. Number of genes as a function of their adjusted P-values. A total of 1397 probes with showing differential expression patterns according to the condition (control vs stress) had P-values equal to or lower than 0.05. The red line shows the threshold (threshold = 0.02) used to select the 31 probes for further analyses.

The initial number of individuals was enough to identify DEGs, but it was not enough to implement BN algorithms. Considering that bioinformatic studies in chickens are not as popular as other models (e.g., rodents or humans), in addition to the complexity of the experimental designs, the following strategy was used to increase the number of observations. The previously identified DEGs were traced to another four datasets that shared the same animal model (chickens), the same tissue (brain - thalamus/hypothalamus), and used the same high-throughput technology to measure the expression of genes (microarrays). Although not all the experiments evaluated the effects of stress (ArrayExpress accession numbers: E-MTAB-924, E-MTAB-3319, E-MTAB-644, and E-MTAB-645), the initial set of 31 probes was used

as the “stress signal”. Consequently, a stress condition could not be included as a variable. The expression values for each gene were extracted after normalizing and correcting the background noise, using the R package “affy”. Thereafter, the 31 probes were annotated, and among the probes, there were 6 sets of 2 probes coding for the same gene (12 probes coding for 6 different genes; 2 “duplicated” probes x 6 genes = 12; 31 probes – 6 “duplicated” probes = 25 genes). The expression values coding for the same gene were averaged into one single value, representing the mean between the two expression values for each pair of observations. Before merging the five datasets into one larger dataset, and considering that each dataset evaluated different experimental designs, each individual dataset was discretized into three categories based on the gene expression values, using a quantile discretization method, to remove potential noise (Balov, 2013). The three categories were low, medium, or high (ordinal; low < medium < high), depending on the gene expression values. The final dataset consisted of 25 genes and 46 individuals, and it was used for further analysis.

4.3. The Bayesian network approach.

BNs were learnt in Banjo, available for free for academic purposes from http://www.cs.duke.edu/~amink/software/banjo/. The algorithm implemented in this study was Simulated Annealing, exploring the search space from an empty graph, and scoring each network with the Bayesian Dirichlet (BDe) score. Banjo allows the possibility to adjust the search parameters; in this particular study, the search space was explored with a total of 250 million networks with local random moves as the proposer. An initial consensus network was built combining the top 100 high-scoring networks provided by Banjo. BN algorithms search the space based on a given set of variables, adding, removing, or reversing edges. After each one of these changes, the new network is scored, and its score is compared to the score corresponding to the previous network: if the score of the new network is higher than the previous network, the latter is discarded, and the process is iterated until no further

improvements in the score metric are made (Heckerman et al., 1995). It is important to bear in mind that the simulated annealing algorithm can accept networks with lower scores as it uses the temperature parameter (Matthäeus et al., 2010). Considering that the process is based on heuristic random searches, different searches had slight differences in the final set of edges (Figure 6).

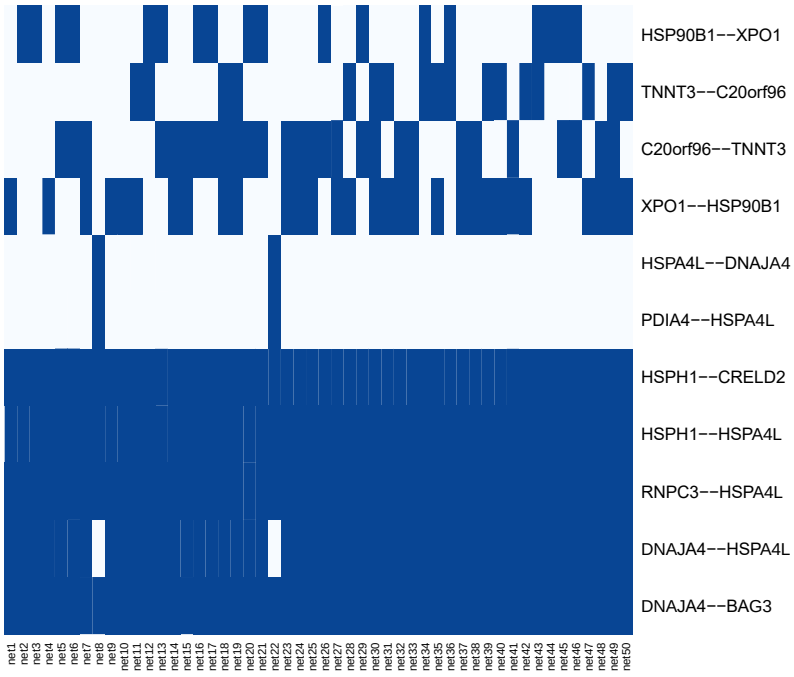

Figure 6. Visualisation of presence or absence of arcs of 50 consensus Bayesian networks learnt in Banjo. The heatmap shows the arcs found by Simulated Annealing while searching the space, visiting a total of 250 million networks. Each column represents a consensus Bayesian network built by combining the top 100 highest scoring networks. Rows represent arcs found by the algorithm. Those arcs present in at least 50% of the networks (threshold = 25 out of 50 networks) were selected to build the weighted network.

Therefore, a weighted network was built by combining the results of 50 consensus networks (each one of them built by combining 100 highest scoring networks) into a matrix of presence/absence: for each individual network, if an edge was present a value of one was assigned, while if the edge was absent, a value of zero was assigned. The values for each edge were added across the 50 networks; the resulting values ranged from 1 to 50. The weighted consensus networks were built using edges present in at least 50% of the networks

(25 times or more). To evaluate the consistency of this approach, it was repeated four times, and there were no differences in the structure of the resulting weighted networks (each weighted network consisted of the same set of edges). In addition to DAVID and the resources accessible in NCBI, the STRING database, a database of known and predicted protein-protein interactions (https://string-db.org), was explored to get further insight into one of the discovered informative relationships. Figure 7 gives the overall view of the steps taken and the decisions made throughout the experimental chapter.

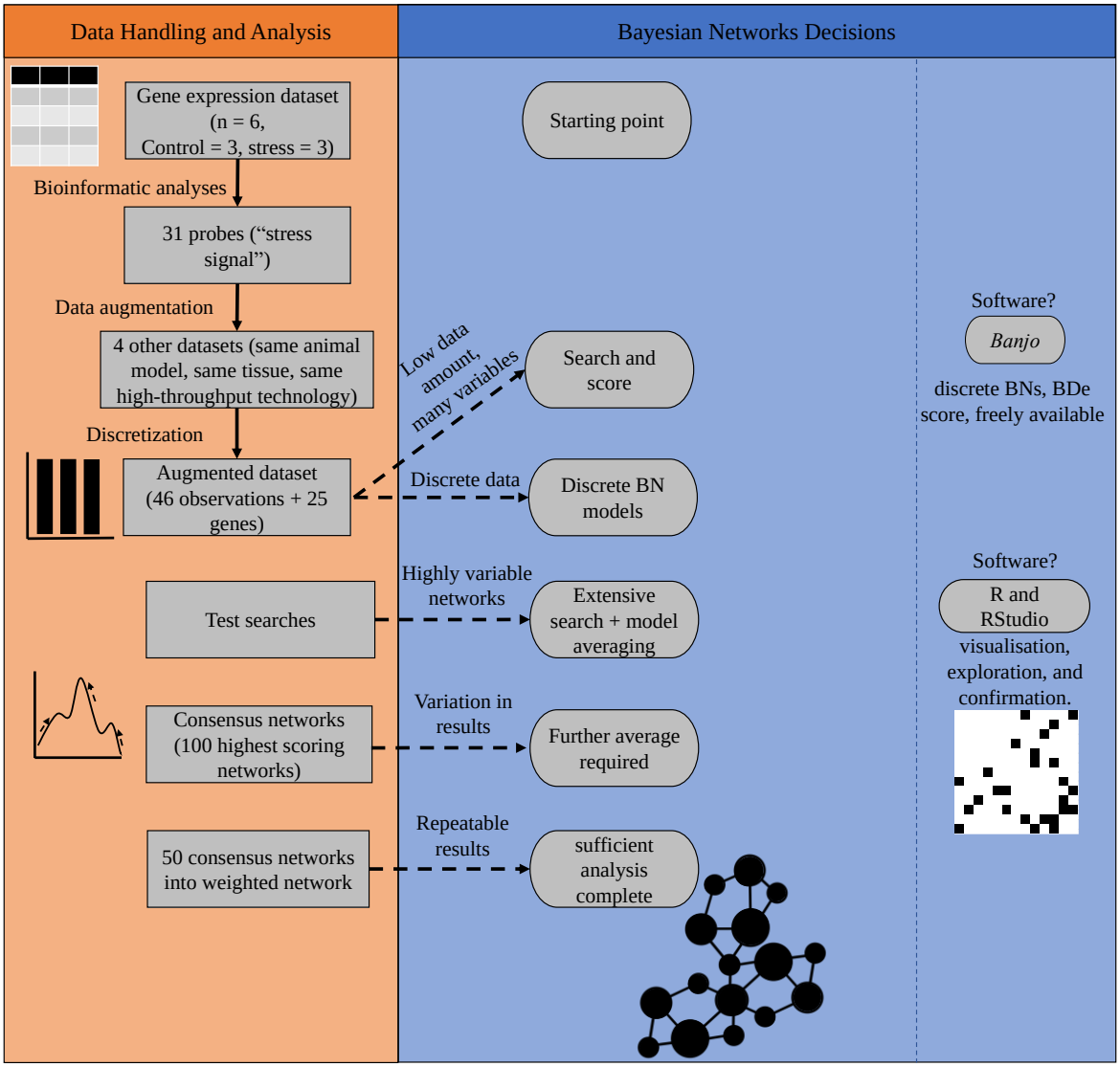

Figure 7. Steps taken and decisions made to build the weighted Bayesian network (BN). The starting point was a dataset consisting of 3 chickens under control and 3 chickens under heat stress conditions. Bioinformatic analyses were performed to normalize, to correct the background noise, and to identify differentially expressed genes (DEG). A total of 31 probes showing differential expression patterns were searched in four other datasets sharing the same animal model (chicken), the same tissue (brain), and the same high-throughput technology (microarray). Each dataset was individually discretized into three-state variables and then merged into a larger dataset consisting of 46 observations and 25 DEGs (12 probes coded for 6 genes (2 "duplicated" probes x 6 genes = 12 probes; 31 probes – 12 "duplicated" probes = 25 DEGs); the corresponding expression values of each one of these "duplicated" probes were averaged into one single value by duplicated probe). The software Banjo was utilised to learn discrete BNs, exploring the search space with a simulated annealing and the BDe score, visiting a total of 250 million networks. An initial consensus BN was built by combining the top 100 highest scoring networks. Heatmaps were used to visualize the results different consensus BNs and due to variation in the final sets of arcs, 50 consensus BN were further combined into a weighted network, by selecting those arcs present in at least 50% of the consensus BNs (threshold: 25 out of 50 networks).

4.4. Results and partial discussion.

The overall structure of the weighted network is shown in Figure 8. The consensus network consisted of only 10 genes out of the initial set of 25 genes with differential expression patterns. Among these genes, 4 heat shock proteins (HSP; *HSPH1*, *HSPA4L*, *DNAJA4*, and *HSP90B1*) were identified as part of the network, interacting not only with each other but also with four other genes. The interaction of these other four genes, *BAG3*, *RNPC3*, *CRELD2*, and *XPO1*, with the HSPs might be closely related to the biological function of these proteins. HSPs are involved in stress tolerance and resistance, playing an important role in protecting the structure of other proteins such as enzymes or receptors, maintaining their functionality (Goel et al., 2021; Perini et al., 2021). During a stress event, especially under the influence of high environmental temperatures, the gene expression levels of these proteins are increased in several tissues such as brain, liver, lungs, heart, and breast (Goel et al., 2021; Perini et al., 2021). It is then plausible that the unravelled interactions discovered by BNs are in agreement with previous studies, with the main difference that, in this study, the relationships were learnt from the data by using a mathematical tool represented by BN algorithms. Two genes, *C20orf96* and *TNNT3*, displayed an interaction between them but did not interact with the HSPs. *C20orf96* is an open reading frame conserved in human, mouse, chicken, and other animal models, that encodes a protein whose function is as yet unknown. As an additional pieces of evidence, the STRING database revealed a protein-protein interaction with Aprataxin (APTX). According to STRING, this protein plays a role in repairing single-strand and double-strand DNA break as well as base excision, sometimes induced by reactive oxygen species. On the other hand, *TNNT3* is a tropomyosin, a member of the tropomyosin family, and it has been reported to play an important role in regulating the growth of dendritic cells in the nervous system of *Drosophila*, in close association with another gene, "flamingo" (Li & Gao, 2003). Based on the exploratory nature of learning the structure of BNs from the data, it is important

to highlight the value of this identified relationship between *C20orf96* and *TNNT3*, as it represents a subject for the development of further research aimed at studying the genetic implications of the functional interactions with APTX and flamingo.

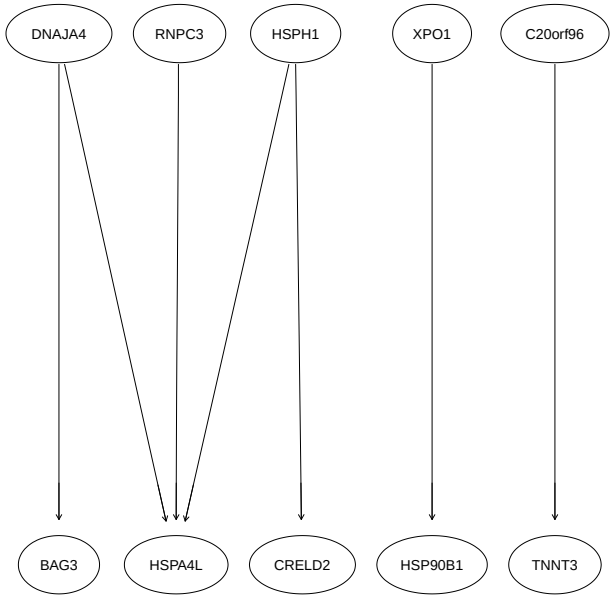

Figure 8. Bayesian network corresponding to highly significant genes related to stress. Nodes correspond to genes, while edges represent the relationship between genes. The network was built considering the edges present in at least 25 out of 50 consensus networks. Note the direction of the arrows do not represent causation, but instead a statistical relationship.

Considering the ten genes included in the network and based on the outcomes of the DAVID bioinformatic tool, two terms were overrepresented within the set of genes: protein processing in endoplasmic reticulum (P = 0.003, Benjamini adjusted P = 0.013) and cytosol (P = 0.006, Benjamini adjusted P = 0.12). Even though HSPs can be found outside the cell, potentially as stress signals, their biological functions are mostly developed inside the cells (Goel et al., 2021; Perini et al., 2021). The interactions discovered by the implementations of a BN approach between HSPs and the other four genes can be closely associated with the role of these proteins protecting the structures, and therefore, maintaining the correct functioning of other cytosolic proteins (Goel et al., 2021; Perini et al., 2021). Considering the individual annotation terms, the four genes in close association with the HSPs had terms such as protein folding and refolding, protein processing in endoplasmic reticulum, endoplasmic reticulum

chaperone complex, among others (Table 1). Taking into consideration the biological functions of HSPs, the unravelled informative relationships and interactions between these four genes and the HSPs might be relevant to the physiology of birds, especially during stress conditions, when maintaining the optimum functioning of the cell machinery would be crucial to deal with the stressor (Goel et al., 2021; Perini et al., 2021).

Initially, and considering that the data came from experiments measuring gene expression in the brain, the functionality and activity of these proteins can be understood as brain specific. However, previous studies suggest that HSP are expressed, either down-regulated or up-regulated, in different tissues in response to stress. Guo et al. (2020) reported that the exposure to stress (addition of CORT to the diet) differentially affected the gene expression of HSP such as HSPA2, HSPA8, HSP90AA1, and HSPH1 (the latter one also discovered in this experimental chapter) (Guo et al., 2020). Xie et al. (2014) also found differences in the expression patterns of two of the most common HSP, HSP70 and HSP90, in the muscle, heart, and liver when laying hens were exposed to acute and chronic heat stress (Xie et al., 2014). Therefore, although the findings are brain specific, it is plausible to think that HSP are playing a role during stress, with particular emphasis on processes such as protein processing in endoplasmic reticulum and maintaining the correct functioning of other proteins under a stress event (Goel et al., 2021; Y. Guo et al., 2020; Perini et al., 2021; Xie et al., 2014).

Table 1 Functional Annotation Table provided by the Database for Annotation, Visualization, and Integrated Discovery (DAVID) corresponding to the Heat Shock Proteins interacting with other four genes. Terms particularly relevant to the stress condition are highlighted in bold. The gene symbol is underlined in parenthesis.

DnaJ (Hsp40) homolog, subfamily A, member 4 (DNAJA4).	
GO TERMS	Response to heat, **protein refolding,** negative regulation of inclusion body assembly, cytosol, membrane, ATP binding.
Heat shock 105kDa/110kDa protein 1 (HSPH1)	
KEGG PATHWAY	**Protein processing in endoplasmic reticulum.**
Heat shock 70kDa protein 4-like (HSPA4L)	
GO TERMS	Cytosol, ATP binding.
KEGG PATHWAY	**Protein processing in endoplasmic reticulum.**

Heat shock protein 90kDa beta (Grp94), member 1 (HSP90B1)	
GO TERMS	response to hypoxia, **protein folding**, intracellular sequestering of iron ion, **response to stress**, ER-associated ubiquitin-dependent protein catabolic process, retrograde protein transport, ER to cytosol, actin rod assembly, negative regulation of apoptotic process, regulation of phosphoprotein phosphatase activity, cellular response to ATP, nucleus, **endoplasmic reticulum**, endoplasmic reticulum lumen, endoplasmic reticulum membrane, cytosol, plasma membrane, focal adhesion, midbody, extracellular matrix, **endoplasmic reticulum chaperone complex**, perinuclear region of cytoplasm, extracellular exosome.
KEGG PATHWAY	**Protein processing in endoplasmic reticulum.**

In bold are highlighted particularly relevant terms to the stress condition.

This study highlights the importance of interdisciplinary approaches to solve complex biological problems, combining genetics, bioinformatics, and mathematics. Initially, bioinformatics was used to identify a set of genes with differential expression patterns, followed by the application of BNs as a machine learning tool to unravel the relationships between a small set of genes, and finally, the implementation of a publicly available online resource for integrated discovery to understand the biological meaning of the unravelled interactions. It is important to emphasize the power of BNs in discovering and unravelling the relationships among a given set of genes from the data themselves (Heckerman et al., 1995). The initial step of learning the structure of the network involved a set of 25 genes, but only 10 of them were part of the overall structure of the network. Interestingly, even though the remaining 15 DEGs showed differential expression patterns driven by the stress condition, they did not show any possible interaction between them. Additionally, within the set of interacting 10 genes, *C20orf96* is a gene whose function is not known, and the application of a BNs approach identified a potential interaction with a protein belonging to the tropomyosin family (Li & Gao, 2003). This fact also highlights that an exploratory approach with BNs can be implemented in knowledge discovery.

Even though the initial dataset used for identifying genes associated with stress had enough individuals to perform bioinformatic analysis with the aim of identifying genes with differential expression patterns driven by the exposure to the stress condition, it represented a small number for learning BNs. However, in our study, we overcame this challenge with a two-step strategy: firstly, the topmost highly significant DEGs were tracked to other datasets, despite of the complexity of their experimental design, with the aim of increasing the number of observations; and secondly, each dataset was discretized into three-state variables before building the final dataset, with the aim of dealing with the potential noise introduced by each particular experimental design (Balov, 2013).

To conclude, a multidisciplinary approach was implemented to reduce the initial number of genes obtained from high-throughput technologies to a small number of DEGs, followed by the discovery of relationships and interactions between the DEGs. The approach involved the combination of bioinformatics, BNs, and the DAVID bioinformatic resource as a database for biological knowledge discovery. The overall results showed that four HSPs have informative relationships unravelled by the BN, not only between themselves but also with four other genes, potentially highlighting their biological functions to protect the structure and to maintain the correct functioning of proteins within the cell. Considering the exploratory nature of our study, future research can be oriented to determine the discovered protein-protein interactions, evaluating the differences between chickens raised under control conditions and chickens raised under stress conditions.

5. RNA-sequencing, spleen, and stress.

5.1. Introduction.

Exploring the brain as the starting point of the stress response resulted in the discovery of Heat Shock Protein as genes potentially crucial to deal with the influence of a stressor, particularly heat stress, when keeping the correct functioning of the cell machinery would be a key factor in this fundamental structure. Once the stressor is perceived and integrated in the brain, the stress response is triggered, involving a cascade of neuroendocrine responses. The end product of the activation of this cascade of responses is the release of glucocorticoids from the adrenal glands, in particular, corticosterone (Dickens & Romero, 2013; Henriksen, Rettenbacher, & Groothuis, 2011; Nazar, Videla, Fernandez, Labaque, & Marin, 2018; Romero, Dickens, & Cyr, 2009; Scanes, 2016; Zulkifli, Al-Aqil, Omar, Sazili, & Rajion, 2009). Corticosterone itself has a major impact on the immune system, and based on this fact, the immune-neuroendocrine interplay is defined as the interactions between three main systems: the nervous, the endocrine, and the immune system.

Considering that corticosterone plays a role in modulating the immune responses such as the recruitment and mobilization of leukocyte populations, modifications of the microbial communities living in the gut, morphological modifications of the structure of the gastro-intestinal tract, and the gene expression patterns, it was worth studying the genetics behind the stress response, and consequently, the stress phenomenon in one of the major immune organs in avian species: the spleen (Calefi et al., 2016; Cantet et al., 2021; Dhabhar, 2009; Elfwing et al., 2015; Løtvedt et al., 2017; Noguera, Aira, Pérez-Losada, Domínguez, & Velando, 2018; Quinteiro-Filho et al., 2010; Vandana et al., 2021; Wickramasuriya et al., 2022). The exposure of chickens to stress causes major changes in the spleen, as a consequence of the effects of glucocorticoids and also by the side effects of the activation of stress response, such as the imbalance between oxidant and antioxidant molecules (Hirakawa et al., 2020; Shini et al., 2008; Van Goor et al., 2017).

Three studies evaluated the effects of stress on the spleen of chickens, measuring the expression patters between the two conditions, in addition to extra factors such as an immune challenge or different breeds (Guo et al., 2020; Park et al., 2019; Van Goor et al., 2017). These studies discovered that there were differences in several biological pathways between non-stress vs stress chickens, such as immune pathways related to proteasome, epithelial adherent junctions and focal adhesion, influenza A, lipid and glycerophospholipid metabolism pathways, protein processing in endoplasmic reticulum, or the regulation of CORT-induced stress effects on immune function (Guo et al., 2020; Park et al., 2019; Van Goor et al., 2017).

These three studies were carried out individually, providing insights into the stress phenomenon in the spleen of chickens. Considering that combining data that have in common the same animal model and the same tissue can provide a more general overview of the same phenomenon as well as further insights into it, a similar approach as the one implemented in the previous chapter was applied. Data coming from two separate experiments were reused and analyzed by BNs in a multi-step interdisciplinary approach, combining the measurements of genetic variables in addition to bioinformatics and mathematics to understand the biology behind a complex biological system. Publicly available repositories were explored to collect data coming from two experiments where the expression values of genes were measured in the spleen of chickens exposed to heat stress by RNA-sequencing. Bioinformatic analyses were implemented to identify a set of genes driven by the exposure to the stressor, followed by learning the structure of a BN to display the relationships and interactions between the genes and the stressful condition. The structure of the network was divided into communities of densely connected nodes, with a special focus on the community of nodes related to the stress condition. Finally, the biological meaning of the discovered interactions and relationships was explored.

5.2. Dataset.

From the previously mentioned studies (Guo et al., 2020; Park et al., 2019; Van Goor et al., 2017), only two datasets were available from a publicly available data repository (Gene Expression Omnibus – GEO), under the following accession numbers: GSE119387 and GSE85434. Briefly, chickens studied in the GSE119387 dataset came from two different regions of Ethiopia: low altitude regions are hot and humid, with chickens adapted to heat conditions, and high-altitude regions, on the other hand, are cooler, with chickens susceptible to heat conditions. The effects of heat stress conditions were evaluated in chickens coming from both regions, but they were raised in low altitude regions (hot and humid). Chickens studied in GSE85434 were exposed either to thermoneutral condition (25°C - control) or to thermal treatment (35°C - heat stress) for 3.5 hours. Therefore, both datasets evaluated the effects of heat stress on gene expression in the spleen of chickens, measured by RNA-sequencing technologies. Each dataset was individually analyzed with the aim of identifying genes relevant for stress, determined by differential expression patterns between non-stress and stress conditions. The .txt files were downloaded, imported into R (R Core Team, 2021), and pre-processed using the R package "edgeR" (Robinson, McCarthy, & Smyth, 2009), normalizing and removing any possible background noise associated with the data. Thereafter, the *lmfit* function, from the R package "limma" (Ritchie et al., 2015), was implemented to fit a linear model according to the experimental design of each dataset. The *eBayes* function was applied to calculate the statistics that would identify the set of genes. Finally, the top highly significant genes were selected using the *topTable* function. A list of 677 and 483 relevant genes were independently identified for each dataset (Figure 9). The possibility that both datasets shared a common signal was examined by looking for genes in common between the two datasets, giving a total of 19 genes shared by the two studies, and representing a manageable number of variables to learn the structure of BNs.

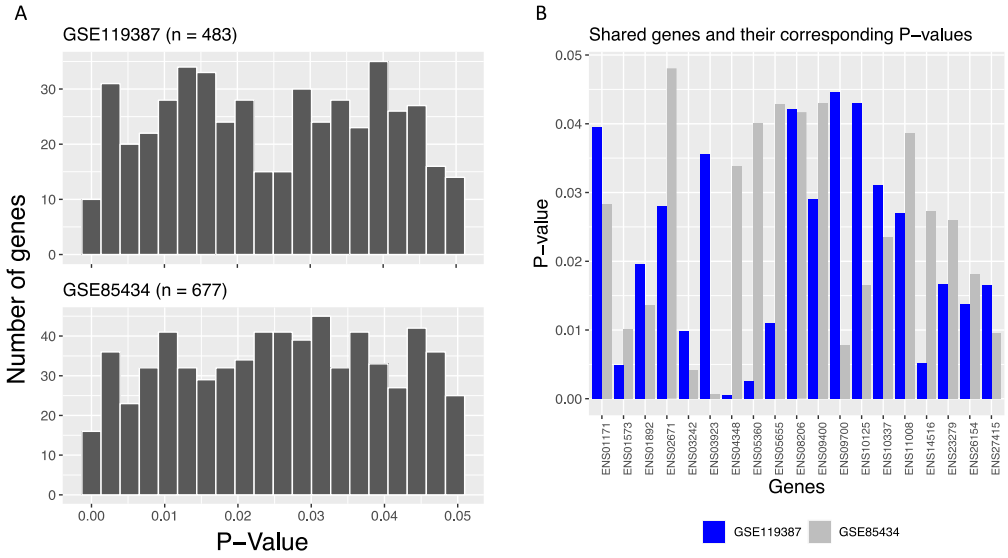

Figure 9. Visualization of P-values per dataset. Number of probes (A) and genes (B) according to their P-values per dataset. Top: GSE119387, bottom: GSE85434; n represents the number of probes showing differential expression patterns (A). Shared genes by the two datasets and their corresponding P-values. Blue bars: GSE119383, grey bars: GSE85434. Genes are represented by their ENSEMBL GENE ID coded as the first three letters and the last five numbers (e.g., ENSGALG00000001573 is coded as ENS01573) (B).

The expression values for each gene were extracted from the datasets and were used to create the final dataset, which consisted of 19 genes and 50 individuals. This dataset was discretized based on the gene expression values into three categories, low, medium or high (ordinal, low < medium < high), applying the function *discretizeDF*, within the "arules" package (Hahsler, Grün, & Hornik, 2005), following a quantile discretization distribution. The discretization of the data was implemented to reduce noise possibly related to differences between experimental designs as well as to increase the statistical power (Heckerman et al., 1995; Milns et al., 2010). Once the data had been discretized, the stress condition was included as a binary variable, taking a value of 0 for chickens raised under non-stress conditions and a value of 1 for chickens exposed to the stress condition.

5.3. The Bayesian network approach.

To learn the structure of BNs, the software Banjo was used (available for free for academic purposes from http://www.cs.duke.edu/~amink/software/banjo/) (Milns et al., 2010;

Mitchell et al., 2021). Banjo implements heuristic searches with two possible algorithms, Greedy or Simulated Annealing and scoring each network with a BDe score. Banjo allows the possibility of selecting the top highest scoring network or combining the n top highest-scoring networks into one consensus network as the output. In this study, the search space was explored with a total of 250 million networks, using a Simulated Annealing algorithm with local random moves as the proposer. A consensus network was built combining the top 100 high-scoring networks. Considering that BNs implement heuristic searches, adding, removing, or reversing edges with the aim of finding the highest scoring network, the final set of edges was slightly different after running the algorithm several times (Figure 10).

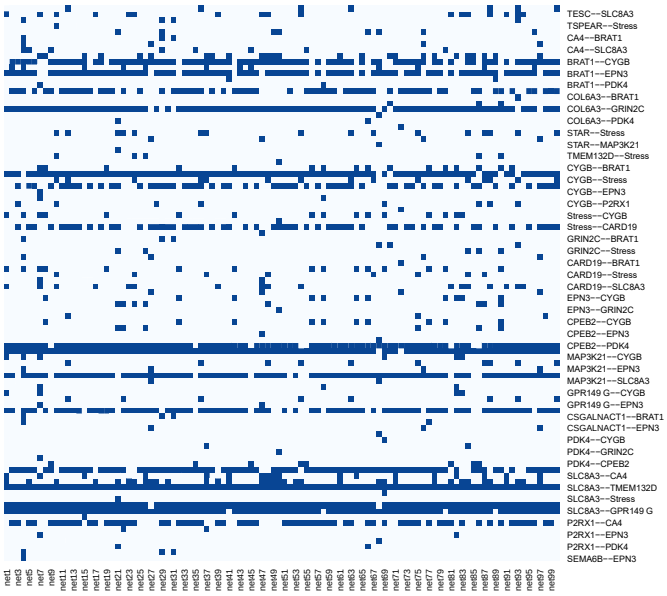

Figure 10. Visualisation of presence or absence of arcs of 100 consensus Bayesian networks learnt in Banjo. The heatmap shows the arcs found by Simulated Annealing while searching the space, visiting a total of 250 million networks. Each column represents a consensus Bayesian network built by combining the top 100 highest scoring networks. Rows represent arcs found by the algorithm. Those arcs present in at least 50% of the networks (threshold = 50 out of 100 networks) were selected to build the weighted network.

Therefore, with the aim of finding the BN that best fitted the data, the following strategy was implemented to solve this challenge. The search space was explored a total of 1000 times, resulting in 1000 consensus networks. These 1000 networks were randomly

divided into 10 groups of a hundred networks (10 x 100 = 1000). Within each set of a hundred networks, all the arcs identified among these 100 networks were used to create a matrix of presence/absence. Thereafter, for each individual network, if an arc was present in the set of arcs, a value of one was assigned; while, if the arcs was absent, a value of zero was assigned. Ten absence/presence matrices were further used for calculating the weight of the arcs: the presence/absence values of each of the arcs (either one or zero) were added across networks; the weight of an arc could take values between 1 and 100. Those arcs whose weight values were equal to or higher than 50 were selected to build the network that best fitted the data. To evaluate the consistency of this approach, the rest of the matrices were used, and there were no differences in the structure of the resulting weighted networks (each weighted network consisted of the same set of edges). Once the consensus BNs were built, a further step was taken using the structure of the networks to implement a community analysis. The aim of this community analysis is to identify clusters of nodes densely connected amongst themselves but scarcely connected with nodes between clusters (Newman & Girvan, 2004). The approach utilised is similar to the one used when performing cluster analysis, with the difference that the input is a network. Initially, the algorithm uses the "edge betweenness" score of the arcs, which is the number of shortest paths that it takes from one or more top vertices (the top node/s of the network structure) to other reachable nodes. Thereafter, it uses a hierarchical divisive clustering, identifying the highest connected nodes between two given variables (arc with the highest "edge betweenness" score) and then, removing this arc for the next step (Newman & Girvan, 2004). The process is repeated several times, until the whole network is divided into smaller communities of densely connected nodes (Newman & Girvan, 2004). The R package "igraph" (Csardi & Nepusz, 2006) was used to identify the communities within the consensus BNs. Initially, the function *cluster_edge_betweenness* was applied to group nodes densely connected, and then the

function *dendPlot* was used to visualize the results. Figure 11 gives the overall view of the steps taken and the decisions made throughout the experimental chapter.

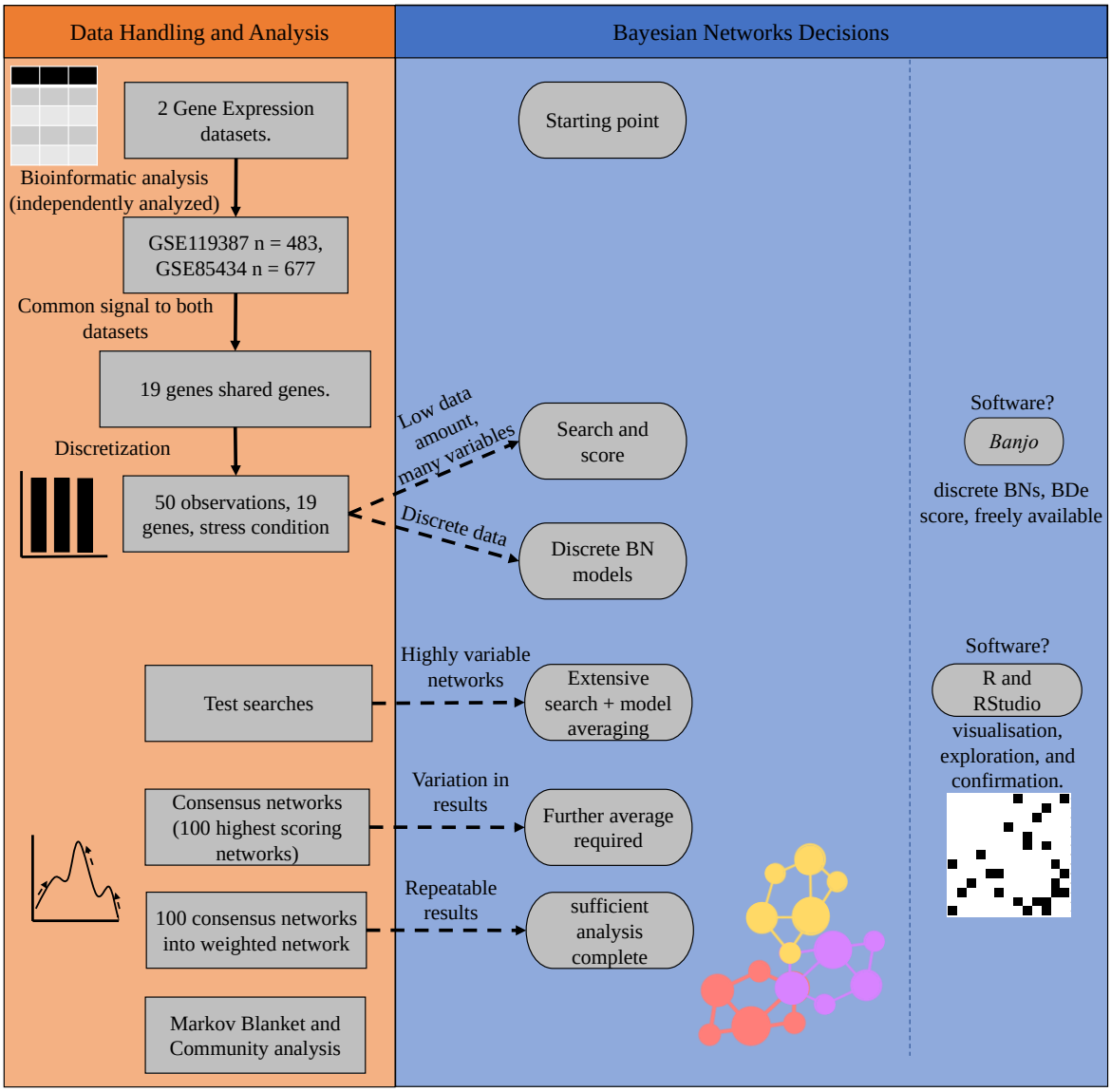

Figure 11. Steps taken and decisions made to build the weighted Bayesian network (BN). The starting point was a set of two studies evaluating stress in the spleen of chickens. Bioinformatic analyses were performed to normalize, to correct the background noise, and to identify genes with differential expression patterns in each dataset. A common "stress signal" was identified between the two datasets: 19 shared genes. The augmented dataset consisted of 50 observations, 19 genes, and as both datasets evaluated stress, the stress condition was included as a binary variable (control = 0; stress = 1). The software Banjo was utilised to learn discrete BN, exploring the search space with a simulated annealing and the BDe score, visiting a total of 250 million networks. An initial consensus BN was built by combining the top 100 highest scoring networks. Heatmaps were used to visualize the results different consensus BNs and due to variation in the final sets of arcs, 100 consensus BN were further combined into a weighted network, by selecting those arcs present in at least 50% of the consensus BNs (threshold: 50 out of 100 networks). Considering that the stress was included as the variable of interest, the Markov Blanket property of BN was applied to identify key genes related to stress. Additionally, communities of densely connected nodes were identified to increase the number of genes closely related to the stress condition.

5.4. Results.

A total of 19 genes having differences in the expression patterns between non-stress and stress chickens were common to two datasets evaluating the effects of stress in the spleen of chickens. The overall structure of the network revealed that 16 out of 19 genes were part of the network in addition to the stress condition (Figure 12). The stress condition was directly connected in the network with only one gene, CARD19. The Markov Blanket property of the condition revealed that in addition to CARD19, CYGB was also related to the stress condition (Figure 5, rectangle-shaped nodes).

The structure of the BN was further explored by dividing the overall structure into smaller communities of densely connected nodes within the community but scarcely connected with nodes in other communities (Newman & Girvan, 2004). The application of a divisive cluster algorithm that uses the structure of the Bayesian networks as the input revealed five communities of densely connected nodes. The community of the stress condition consisted of four genes: in addition to the genes belonging to the Markov Blanket, BRAT1 and EPN3 displayed a possible interaction with the stress condition, representing a group of genes densely connected amongst themselves, but scarcely connected with the rest of the genes (Figure 5, nodes highlighted in pink).

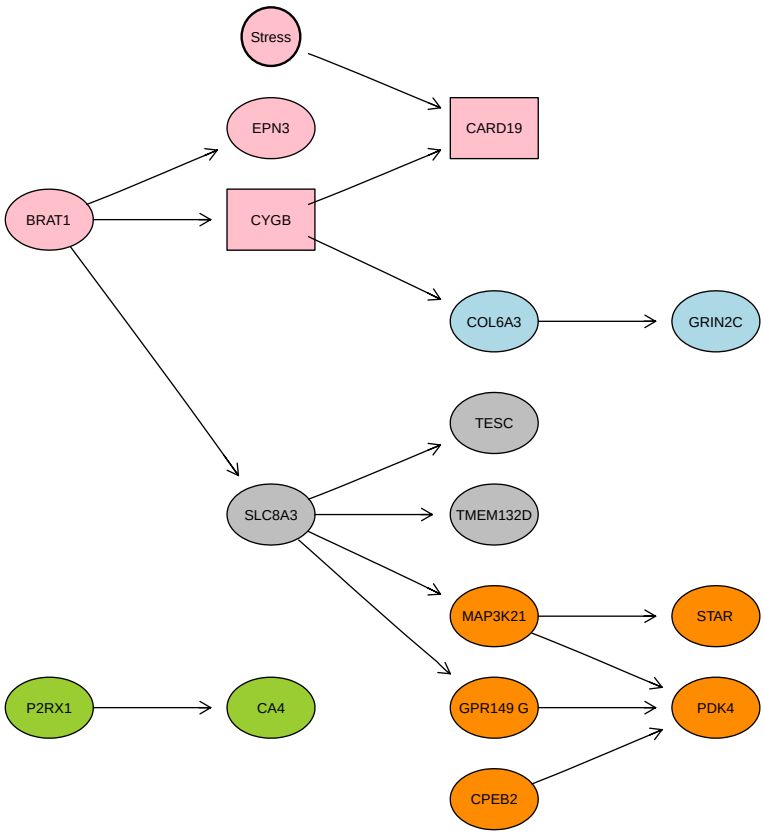

Figure 12. Bayesian network and community analysis of a set of genes. 19 genes showing differences in expression pattern were initially included at the time of learning the structure of the network in addition to the stressful condition; however, only 16 out of those 19 were linked in a network structure. Nodes represent each one of the genes and the stressful condition (circle-shaped node, thick outline), the edges represent probabilistic dependencies between the nodes. Note the direction of the arrows do not represent causation, but instead a statistical relationship. The Markov Blanket of the stress condition (rectangle-shaped nodes) consisted of two genes, CARD19 (child) and CYGB (spouse). Five communities of densely connected nodes were identified (different colours represent different communities). The community of the condition consisted of 4 genes (CARD19, EPN3, CYGB, and BRAT1, highlighted in pink).

DAVID was applied to explore the biological functionality of these four genes. DAVID overrepresentation analysis identified three terms, although all were not significant after adjustments for multiple tests: Calcium signaling pathway (KEGG Pathway, P-value = 0.026, Benjamini adjusted P-value = 0.28), sarcolemma (GO term, P-value = 0.045, Benjamini adjusted P-value = 1), and membrane (keyword, P-value = 0.06, Benjamini adjusted P-value = 1). The DAVID Functional Annotation Table (Table 2) shows the KEGG pathways (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012) and GO TERMS (Ashburner et al., 2000) for each one of the genes being part of the community analysis of the stress condition. Terms

particularly relevant to the stress condition are related to regulation of apoptotic process and caspase recruitment (CARD19), oxygen transporter activity and oxygen binding (CYGB), apoptosis process and cellular response to DNA damage stimulus (BRAT1), and endocytosis (EPN3).

Table 2. Functional Annotation Table provided by the Database for Annotation, Visualization, and Integrated Discovery (DAVID) corresponding to the four genes found to be in close relationship with the stressful condition. Terms particularly relevant to the stress condition are highlighted in bold. The gene symbol is underlined in parenthesis.

Caspase recruitment domain family member 19 (CARD19)	
GO TERMS	**Regulation of apoptotic process**. Integral component of membrane.
Cytoglobin (CYGB)	
GO TERMS	Neuron projection, neuronal cell body. **Oxygen transporter activity**, iron ion binding, **oxygen binding**, heme binding.
BRCA1 associated ATM activator 1 (BRAT1)	
GO TERMS	Positive regulation of protein phosphorylation, glucose metabolic process, apoptotic process**, cellular response to DNA damage stimulus**, cell proliferation, response to ionizing radiation, cell growth, cell migration, mitochondrion localization. Nucleus, cytoplasm, membrane.
Epsin 3 (EPN3)	
KEGG PATHWAY	**Endocytosis**.

In bold are highlighted particularly relevant terms to the stress condition.

5.5. Partial discussions.

This study was aimed at identifying a reduced number of genes closely associated with a stressful condition in the chicken as a poultry animal model. To have a more accurate approximation to the stress phenomenon, two publicly available datasets involving the measurement of gene expression in the spleen of chicken exposed to heat stress were combined into a larger dataset. After bioinformatic pre-processing and analysis, a set of 19 genes common to both datasets with differential expression patterns was identified; these genes were used for learning the structure of the BNs. With the BN in place, its structure was divided into smaller communities of densely connected nodes. By the implementation of this approach, two genes

were identified as part of the Markov Blanket property of the stress condition. In addition to these two genes, two other genes were part of the community of the condition, giving a total of 4 out of the 19 initial genes displaying a close relationship with the stress condition. The results showed a small set of relevant genes related to stress that can be used to extract meaningful information regarding the genetics of this complex phenomenon.

Stress involves the perception of the stimulus in the immune-neuroendocrine interplay, triggering the stress response, and displaying physiological and behavioral adaptations with the aim of dealing with the stressful stimulus (Ashley & Demas, 2017; Calefi et al., 2017; Sapolsky, Romero, & Munck, 2000; Selye, 1950). Heat stress has been widely studied and its effects on immune organs and immune responses have been reported (Calefi et al., 2017; Honda et al., 2015; Mashaly et al., 2004; Quinteiro-Filho et al., 2012). In particular, Hirakawa et al. (2020) found that the mass of the spleen was severely affected by the exposure to heat stress (Hirakawa et al., 2020). Additionally, heat stress altered the structure of the spleen, having an impact on the humoral immune responses that modulate the lymphocyte populations (Hirakawa et al., 2020). Chickens under high environmental temperatures have also shown imbalances in the oxidant/anti-oxidant status as a consequence of the alteration of some by-products or end products of lipid peroxidation such as malondialdehyde (MDA) and thiobarbituric acid reacting substances (TBARS) (Altan, Pabuçcuoğlu, Altan, Konyalioğlu, & Bayraktar, 2003; Lin, Decuypere, & Buyse, 2006). The imbalance is created by the excess of oxidant molecules, such as reactive species containing oxygen, nitrogen, and/or chlorine, potentially affecting the structure of proteins, lipids, and DNA and RNA. Consequently, the functioning of the cell might be affected in terms of energy availability, calcium homeostasis, and mitochondrial functionality, leading to cell damage, and therefore to the survival of the cell being threatened by apoptosis or necrosis (Akbarian et al., 2016; Chen, Ning, Zhang, Tang, & Teng, 2020).

The Markov Blanket property together with the community analysis revealed a total of four genes in close association with the stress condition. One of the genes, CARD19, showed a direct interaction with the stress condition, while the other three genes were part of the Markov Blanket and/or the community of the stress condition. CARD proteins belong to the family of caspase recruitment domains and they are proteins that mediate apoptosis as well as the activation of the NF- κβ signaling pathways (Bertin et al., 2000, 2001; Hofmann, 1999; Wang et al., 2001). Cytoglobin (CYGB) belongs to the globin family, whose major role is related to the provision of oxygen in different tissues and organs, in addition to a potential protective activity against reactive oxygen species (Burmester, Ebner, Weich, & Hankeln, 2002; Kugelstadt, Haberkamp, Hankeln, & Burmester, 2004; Schmidt et al., 2004). Considering that heat stress leads to oxidative stress, cell damage, apoptosis, and immune dysfunction, CARD19 and CYGB could be identified as key genes associated with these mechanisms that chickens trigger as a consequence of the influence of the stressor. Under exposure to other stressors, such as an immune challenge with mycotoxin or hypoxic conditions, apoptotic signaling pathways were also activated in splenic cell (Chen et al., 2020; Ren et al., 2015). Specially under hypoxic conditions, Chen et al. (2020) identified that splenic cells initiated apoptotic signaling pathways as a result of oxidative stress involving inflammatory mechanisms and the NF- κβ pathway (Chen et al., 2020).

Considering the learnt structure of the BN, the further analysis of smaller communities of densely connected nodes showed that the stress condition potentially interacted with two other genes: BRAT1 and EPN3. BRCA1-associated ATM activator 1 (BRAT1) was previously identified by Qui et al. (2018) in the spleen of layer chickens undergoing an infection with avian leukosis virus (subgroup J) (Qiu et al., 2018). In humans, this gene interacts with two other genes, BRCA1 and ATM, mediating cell pathways associated with DNA damage as well as apoptosis (Aglipay et al., 2003; Aglipay, Martin, Tawara, Lee, & Ouchi, 2006; Okada &

Ouchi, 2003; Scully et al., 1997). Epsin-3 (EPN3) is a member of the endocytosis protein adapter gene family, and its main function is related to endocytosis (Xie, Cho, & Fischer, 2012). Additionally, EPN3 has been identified in pathological or damaged tissues requiring wound healing (Spradling, McDaniel, Lohi, & Pilcher, 2001). All in all, it seems that both the Markov Blanket and the community of densely connected nodes of the stress condition are pointing towards key genes related to apoptosis and tissue damage. It is then plausible to highlight that when chickens are exposed to a complex phenomenon such as heat stress, one of the main immune organs, the spleen, reflects some morphological and physiological alterations as a consequence of undergoing apoptotic-related mechanisms, potentially translating into the reported suppression and dysfunction of the immune responses (Guo et al., 2020; Hirakawa et al., 2020).

Regarding the network approach implemented in this study, the combination of two strategies was applied with the aim of identifying genes in close relationship with the stress condition: the Markov Blanket property of Bayesian networks and the community of highly connected nodes (Aliferis et al., 2010; Newman & Girvan, 2004). Initially, the overall structure of the network, the relationships, and the interactions between the given set of variables (the genes and the stressful condition) were learnt from the data. As a following step, the already learnt structure of the Bayesian network was divided into smaller groups of densely connected nodes (Newman & Girvan, 2004). By combining these two strategies, an initial set of 19 genes were further reduced to a small set of genes that showed a close association with the stress and that can be further studied. Among this small set of genes, the structure of the BN revealed that CARD19 showed a close interaction with the stressful condition, suggesting this gene could be explored as a potential biomarker of stress. Therefore, further research can be developed with short-term goals, such as using these genes to identify chickens raised under non-stress or stress conditions, and consequently, using them as indicators of stress, raising the alarm to monitor

and manage the breeding conditions to mitigate the detrimental effects of stress on poultry production (Lara & Rostagno, 2013; Renaudeau et al., 2012). On the other hand, further research can have long-term goals, such as artificial selection and breeding programs in order to enhance the resilience or resistance of chicken breeds to stress, such as the Fayoumi chickens that have been used in studies as a heat stress and disease resistant breed (Redmond, Chuammitri, Andreasen, Palić, & Lamont, 2009; Renaudeau et al., 2012; Van Goor et al., 2017).

In conclusion, this study implemented a series of steps aimed at reducing an initial number of genes obtained from high-throughput technologies to a small number of genes and unravelling their informative (functional) relationships and interactions. Two previous studies that evaluated the effects of stress on the spleen of chickens were combined to get a more accurate approximation to the stress phenomenon. The series of steps involved the combination of: i) bioinformatic tools to identify differentially expressed genes, ii) BNs to learn the overall structure of the network, iii) the Markov Blanket together with the community analysis to identify a small set of genes in close association with the stress condition, and iv) the database for biological knowledge discovery DAVID. Such a sequence of computational approaches could be applicable to many studies of gene expression, across many measurement platforms, enabling combination of power from multiple experiments to identify of small sets of genes for further study. Previous studies showed that heat stress has an impact on gene expression in the spleen (Guo et al., 2020; Park et al., 2019; Van Goor et al., 2017). Park and collaborators (2019) found that chickens susceptible to high environmental temperatures might also be susceptible to immune challenges, considering that immune pathways such as focal adhesion, influenza A, or the signalling pathway associated with the family of receptor tyrosine kinases (ErbB signalling pathway) were enriched when birds were exposed to heat stress (Park et al., 2019). Van Goor and collaborators (2017) found that gene expression was upregulated when

chickens were exposed to heat stress, and genes were related to intestinal permeability, leading to an alteration of the intestinal mucosa and its functionality (Van Goor et al., 2017). Guo and collaborators (2020) found pathways related to immunity and endoplasmic protein processing. Among the genes with differential expression patterns, some Heat Shock Proteins, such as HSPA2 and HSPH1, as well as some cytokines were found (Gou et al., 2020). The work presented in this experimental chapter is complementary to the original studies, in that the identified genes are not an overall picture (this would be found in those studies), but instead provide information about a small set of genes with a strong signal, across multiple studies, suggesting relation to the condition of interest. Here, the outcome of this series of steps identified two genes as being part of the Markov Blanket and two additional genes as being part of the community analysis for the stress condition in poultry. The biological processes of these four genes were related to damage and apoptosis, and they could potentially be further used as biomarkers of heat stress. The exploratory nature of this study requires future research to determine whether the genes can potentially be used as hallmark genes when comparing chickens raised under non-stress conditions and chickens raised under stress conditions. Additionally, as the BN unravelled some informative interactions between genes that belonged to the community of the stress condition, such as BRAT1-EPN3, BRAT1 - CYGB, and CYCG-CARD19, it could also be possible to study more in-depth these protein-protein interactions in biological samples, and consequently, evaluate possible epistatic interactions.

6. Epigenetics, brain, and stress.

6.1. Introduction.

In previous experimental chapters, the expression values of genes were measured by two high-throughput technologies: microarray and RNA-sequencing. Measuring expression values involves the mRNA as the starting point, that is the sequence of RNA that will code for a particular protein, once the fragment is properly pre-processed and transformed into the corresponding protein by cell machinery. Considering the stress phenomenon, throughout the introduction and the previous experimental chapters, bioinformatic tools have been implemented to identify differences in the expression patterns between non-stressed and stressed birds (Fallahsharoudi et al., 2017; Guo et al., 2020; Luo, Song, Ji, Zhang, & Zhang, 2014). However, previous results demonstrate that there are two main drawbacks of implementing only bioinformatic tools with that particular aim: i) the list of genes can be quite extensive, which can difficult the search of the biological meaning of the genes; and ii) being part of the list of genes with differential expression patterns does not mean that they will display informative relationships and interactions with other genes. Additionally, some of the genes that were part of the consensus BNs did not have a biological function yet discovered, representing the possibility to further exploring these genes in future studies.

When working in the genetic field, measuring the expression values of genes is not the only mechanism and/or process that can be evaluated under two experimental conditions. There are several genetic changes that can be measured, ranging from single nucleotide polymorphisms (SNP) to epigenetic modifications of the DNA sequence, such as the addition of methyl groups or chromatin folding (Perini et al., 2021; Pértille et al., 2017, 2020; Skinner et al., 2010). In the case of methylation, the addition of methyl groups to the DNA sequence as an epigenetic modification, can have an impact on gene expression (Pértille et al., 2017, 2020; Skinner et al., 2010). Taking this into consideration, it was worth exploring this field of research, on the one hand, to tackle the stress phenomenon from another point of view, that can

be seen as a broader view of the same phenomenon, considering that it implies an external modification of the DNA, and on the other hand, to gain a better understanding and further insights into the stress phenomenon itself.

In this chapter, the aim of the study was two-fold: first, and in line with the previous two chapters, to apply a BNs approach in order to provide further insights into the relationships between epigenetics and induced stress in a poultry animal model, the chicken (*Gallus gallus*). Second, working with the corresponding dataset was complex, especially when learning the consensus network. It required a series of steps that started from 100 random graphs, applying a BN algorithm to each individual random graph, creating presence/absence matrices including the individual scores of the networks, the application of an ecological average method, and ended with the selection of arcs present in at least 50% of several searches. Therefore, the second aim was to clearly lay out the decision-making process in order to provide a roadmap to enable others to make principled choices when undertaking BNs analysis.

6.2. Dataset.

The data was accessed and downloaded from the European Nucleotide Archive (ENA, www.ebi.ac.uk), under the accession number PRJEB34868 (Pértille et al., 2020). The dataset consisted of 46 male White Leghorn chickens (*Gallus gallus*). The experiment involved 0-26 days aged chickens, 22 raised under control conditions, while the other 24 were exposed to a social isolation protocol. This isolation protocol was applied from the day 4 of age until the day 26 of age (period of 21 continuous days), as described by Pértille et al. (2020). Briefly, birds under the stressful condition were daily exposed to social isolation for one hour during the first week, two hours during the second week, and three hours during the third and final week. During the exposure to the isolation stress, birds were individually placed in a box with vocal but no visual or physical contact with other birds. Thus, during the stress treatment, birds were exposed to a combination of stressors: social isolation and deprivation of food and water

(Goerlich et al., 2012; Pértille et al., 2020). The control animals were not exposed to the social isolation protocol, but they were raised under the same environmental conditions as the stressed birds. The identification of differentially methylated regions (DMRs) between these experimental groups included a series of steps such as blood collection at day 26 of age (2 hours after the last day of isolation was ended) in order to extract the DNA from red blood cells, the preparation of the libraries using the GBS-MeDIP method (Rezaei et al., 2022) to sequence the DNA fragments and finally the bioinformatic pre-processing and analysis to identify the DMRs (Pértille et al., 2017, 2020). The DMRs identified in this study were selected by first defining 'Regions of Interest' (ROI) showing differences in sequencing coverage between the treatment and control groups. This was done with MACS2, which is a recommended tool to identify sample-wise 'peak specific' methylated regions of variable sizes in experiments using paired controls to determine enrichment against background (Cavalcante, Qin, & Sartor, 2018; Feng, Liu, Qin, Zhang, & Liu, 2012; Niazi, Geyer, Vickers, Hoffmann, & Swain, 2016). Then, we applied the weighted trimmed mean of M-values (TMM) method within "edgeR" on these ROI obtained with MACS2. TMM is used to calculate scale factors between libraries. One of the standard outputs of this "edgeR" test is a p-value (*edgeR.p.value*). Based on this, 60 DMRs were selected with $p \leq 0.005$. DMRs were annotated and divided into 4 different categories based on the features of the genome in the region: promoter, distal intergenic, intron, or exon following the methods described by Pértille et al. (2020). Briefly, the categorization is based on the position of the DMR on the chromosome and the distance between the DMR and its nearest transcription starting site (Pértille et al., 2020). DMRs categorised as promoters, introns, and exons were annotated with the corresponding gene name. Promoters, introns, and exons without a gene symbol were assigned their corresponding ENSEMBL gene name using the first three letters and the numbers after the zeros (e.g. ENS50641 represents ENSGALG00000050641.1). DMRs annotated as DMR1 up to DRM7

correspond to distal intergenic regions without a gene symbol. A list of the 60 DMRs used and their annotations is provided in Table 3.

Table 3. Differentially methylated regions and their annotations. List of differentially methylated regions (DMR) with their corresponding genetic annotation terms. The first column ("SYMBOL") represents the abbreviated gene name of the methylated region; those which say "annotated" plus a number means that the symbol for that particular DMR was not available; the second column ("Gene ID") represents the ENSEMBL gene ID; the third column ("Description") represents the description of the DMR (NA for those not available); the fourth column ("Type of DMR") represents the type of DMR (e.g., Promoter, Intron, etc.); and finally, the fifth column ("SYMBOL network") represents the name used in the consensus Bayesian network.

DMR	SYMBOL	geneId	Description	Type of DMR	SYMBOL network
DMR 1	annotated 1	ENSGALG00000050641.1	NA	Promoter (<=1kb)	ENS50641
DMR 2	annotated 2	ENSGALG00000035970.3	NA	Promoter (9-10kb)	ENS35970
DMR 3	annotated 3	ENSGALG00000052484.1	NA	Distal Intergenic	DMR1
DMR 4	annotated 4	ENSGALG00000053301.1	NA	Distal Intergenic	DMR2
DMR 5	SHISA2	ENSGALG00000043035.3	shisa family member 2	Promoter	SHISA2
DMR 6	FBN1	ENSGALG00000004960.6	fibrillin 1	Distal Intergenic	FBN1
DMR 7	GNAO1	ENSGALG00000003163.6	G protein subunit alpha o1	Intron (ENSGALT00000005006.5/ENSGALG00000003163.6)	GNAO1
DMR 8	annotated 5	ENSGALG00000053123.1	NA	Distal Intergenic	DMR3
DMR 9	annotated 6	ENSGALG00000051236.1	NA	Intron (ENSGALT00000092657.1/ENSGALG00000051236.1)	ENS051236
DMR 10	annotated 7	ENSGALG00000048708.1	NA	Distal Intergenic	DMR4
DMR 11	VGLL4	ENSGALG00000004937.6	vestigial like family member 4	Promoter	VGLL4
DMR 12	annotated 8	ENSGALG00000041405.3	NA	Promoter	ENS41405
DMR 13	annotated 9	ENSGALG00000047523.1	NA	Distal Intergenic	DMR5
DMR 14	CANX	ENSGALG00000032148.2	calnexin	Promoter	CANX
DMR 15	ARHGAP26	ENSGALG00000033938.3	Rho GTPase activating protein 26	Intron (ENSGALT00000070058.3/ENSGALG00000033938.3)	ARHGAP26
DMR 16	annotated 10	ENSGALG00000044794.2	NA	Distal Intergenic	DMR6
DMR 17	MYH11	ENSGALG00000006520.7	myosin, heavy chain 11, smooth muscle	Promoter	MYH11
DMR 18	TPST2	ENSGALG00000005626.6	tyrosylprotein sulfotransferase 2	Promoter	TPST2
DMR 19	POP5	ENSGALG00000007124.7	POP5 homolog, ribonuclease P/MRP subunit	Promoter	POP5

DMR 20	annotated 11	ENSGALG00000027231.4	NA	Promoter	ENS27231
DMR 21	GPR141	ENSGALG00000032363.2	G protein-coupled receptor 141	Promoter	GPR141
DMR 22	EEPD1	ENSGALG00000039983.2	endonuclease/exonuclease/phosphatase family domain containing 1	Promoter	EEPD1
DMR 23	BOP 1.00	ENSGALG00000053353.1	block of proliferation 1	Exon (ENSGALT00000105884.1/ENSGALG00000053353.1)	BOP 1.00
DMR 24	annotated 12	ENSGALG00000046425.2	NA	Promoter	ENS46425
DMR 25	TTLL9	ENSGALG00000006460.6	tubulin tyrosine ligase like 9	Promoter	TTLL9
DMR 26	ZBTB48	ENSGALG00000000637.4	zinc finger and BTB domain containing 48	Promoter	ZBTB48
DMR 27	DOCK5	ENSGALG00000000311.6	dedicator of cytokinesis 5	Intron (ENSGALT00000000413.5/ENSGALG00000000311.6)	DOCK5
DMR 28	ZDHHC18	ENSGALG00000034910.2	zinc finger, DHHC-type containing 18	Promoter	ZDHHC18
DMR 29	AGO1	ENSGALG00000002249.6	argonaute 1, RISC catalytic component	Promoter	AGO1
DMR 30	RP1-27O5.3	ENSGALG00000003408.3	zinc finger and BTB domain containing 8B	Promoter	RP1-27O5.3
DMR 31	CDK18	ENSGALG00000000692.6	cyclin dependent kinase 18	Promoter	CDK18
DMR 32	MFSD4A	ENSGALG00000000695.6	major facilitator superfamily domain containing 4A	Promoter	MFSD4A
DMR 33	PLXNA2	ENSGALG00000001264.6	plexin A2	Intron (ENSGALT00000001931.6/ENSGALG00000001264.6)	PLXNA2
DMR 34	LOC770074	ENSGALG00000050600.1	uncharacterized LOC770074	Promoter	LOC770074
DMR 35	annotated 13	ENSGALG00000053725.1	NA	Exon (ENSGALT00000100075.1/ENSGALG00000053725.1)	ENS53725
DMR 36	CWC25	ENSGALG00000001624.7	CWC25 spliceosome associated protein homolog	Promoter	CWC25
DMR 37	STAT3	ENSGALG00000003267.7	signal transducer and activator of transcription 3	Promoter	STAT3
DMR 38	OCLN	ENSGALG00000037316.2	occludin	Promoter	OCLN
DMR 39	NRXN1	ENSGALG00000009107.6	neurexin 1	Intron (ENSGALT00000057931.2/ENSGALG00000009107.6)	NRXN1
DMR 40	PCSK2	ENSGALG00000008734.6	proprotein convertase subtilisin/kexin type 2	Promoter	PCSK2
DMR 41	PARK2	ENSGALG00000011562.6	parkin RBR E3 ubiquitin protein ligase	Intron (ENSGALT00000101286.1/ENSGALG00000011562.6)	PARK2
DMR 42	SELENOI	ENSGALG00000016560.6	selenoprotein I	Intron (ENSGALT00000026723.6/ENSGALG00000016560.6)	SELENOI

DMR 43	EFR3B	ENSGALG00000016605.6	EFR3 homolog B	Promoter	EFR3B
DMR 44	MIP	ENSGALG00000042119.2	major intrinsic protein of lens fiber	Promoter	MIP
DMR 45	DOCK11	ENSGALG00000006017.6	dedicator of cytokinesis 11	Intron (ENSGALT00000074876.2/ENSGALG00000006017.6)	DOCK11
DMR 46	TRMT10A	ENSGALG00000012264.6	tRNA methyltransferase 10A	Promoter	TRMT10A
DMR 47	LRP5	ENSGALG00000029533.2	LDL receptor related protein 5	Distal Intergenic	LRP5
DMR 48	annotated 14	ENSGALG00000047746.1	NA	Intron (ENSGALT00000107381.1/ENSGALG00000047746.1)	ENS47746
DMR 49	annotated 15	ENSGALG00000054193.1	NA	Promoter	ENS54193
DMR 50	LOC101750642	ENSGALG00000049221.1	uncharacterized LOC101750642	Distal Intergenic	LOC101750642
DMR 51	LOC107054063	ENSGALG00000046679.1	uncharacterized LOC107054063	Promoter	LOC107054063
DMR 52	DGKD	ENSGALG00000001730.6	diacylglycerol kinase delta	Intron (ENSGALT00000002654.6/ENSGALG00000001730.6)	DGKD
DMR 53	SCHIP1	ENSGALG00000039468.2	schwannomin interacting protein 1	Promoter	SCHIP1
DMR 54	annotated 16	ENSGALG00000044085.2	NA	Promoter	ENS44085
DMR 55	annotated 17	ENSGALG00000010218.6	NA	Promoter	ENS10218
DMR 56	SKOR2	ENSGALG00000051844.1	SKI family transcriptional corepressor 2	Promoter	SKOR2
DMR 57	ARHGEF28	ENSGALG00000014923.6	Rho guanine nucleotide exchange factor 28	Intron (ENSGALT00000024069.6/ENSGALG00000014923.6)	ARHGEF28
DMR 58	annotated 18	ENSGALG00000050012.1	NA	Intron (ENSGALT00000107859.1/ENSGALG00000050012.1)	ENS50012
DMR 59	XRCC4	ENSGALG00000015620.6	X-ray repair cross complementing 4	Intron (ENSGALT00000025179.6/ENSGALG00000015620.6)	XRCC4
DMR 60	annotated 19	ENSGALG00000046715.1	NA	Distal Intergenic	DMR7

6.3. The Bayesian network approach.

6.3.1 Data discretization and contingency test

The DMR dataset (46 samples and 60 variables) consisted of individual counts obtained within the experimentally obtained DMRs described above, corresponding to the number of segments aligned to a particular DNA region, values ranging from 0 to 39. This count data was furthered discretised with the aim of filtering noise as well as increasing the statistical power (Yu et al., 2004). The most statistical power is provided by all discrete states having roughly equivalent numbers of data points (Heckerman et al., 1995; Milns et al., 2010); here, zero counts was the most abundant observation, and thus the closest to this idea was a binary dataset with two categories: zero and one. All original values equal to zero were assigned a new value of zero (no methylation), while the rest of the values were assigned a new value of one (methylation). In addition to the DMRs, the stressful condition was included in the dataset as a binary variable, considering the control condition as 0 and the stress condition as 1 (22 individuals = 0, 24 individuals = 1). The DMRs plus the stressful condition are the features included as nodes in the network. An overabundance of the discrete state of zero remained, thus pair-wise contingency-test filtering was applied as in Milns et al (2020): a chi-square test was applied to all possible pairs of variables, and those with p-values equal or greater than 0.25 were identified as showing no possible dependence between them (Milns et al., 2010). These were included in the BN analysis as a list of arcs to be blocked, representing prior information that these links should be excluded from the network (Nagarajan et al., 2013a, 2013b).

6.3.2 Bayesian network analysis

To explore other publicly available software to learn BNs, in this experimental chapter the R and RStudio were applied. The R package "bnlearn" (Scutari, 2010) was used to learn the structure of the network. Initial tests were done by starting groups of 100 searches from

random graphs generated by the *random.graph* function, using *tabu* search function, with the BDe score and the list of arcs to be blocked included (Nagarajan et al., 2013a, 2013b). Summary networks of arcs found across these groups of searches were analysed for arc correspondence and showed high variability. Variability in search results, e.g., the difficulty in finding the same consensus network after several runs of the algorithm, was confirmed using BayesPiles, a visual analytical system (Vogogias et al., 2018). Complex search spaces, as mentioned in the methods section, require further visual exploration when it comes to the identification of the consensus network. BayesPiles, a visualisation support for BN structure learning, requires the implementation of the Banjo software (Smith, Yu, Smulders, Hartemink & Jarvis, 2006): equivalent settings using the BDe score and a greedy search (closest available to the *tabu* function) were set in Banjo, and the list of arcs to be blocked included. BayesPiles uses the outcomes of Banjo to explore and compare the top scoring networks discovered by Banjo in terms of arcs, scores, and structures (Vogogias et al., 2018). Additionally, arcs can be filtered according to the number of times they are present across the set of top scoring networks (Vogogias et al., 2018). For this particular search space, four sets of searches including multiple starts from random networks were visualised (Figure 7), revealing again high variability.

Thus, a method previously applied in an ecological system with a similarly high variability in search results was used (Milns et al., 2010). This method collects top networks from multiple searches (100 searches both in (Milns et al., 2010) and here), then applies a model averaging approach considering the score of the network to develop probabilities of arcs being in a high-scoring network. These probabilities are clustered into higher and lower probability clusters and are provided uncertainty values for cluster membership. Those arcs in the higher probability cluster (with a probability and uncertainty cut-off) are presented as the final network. To perform this analysis, 100 searches were started from random graphs generated by the *random.graph* function, using *tabu* search function, with the BDe score and

the list of blocked arcs included, as above, identifying 100 top networks. The arcs present in the 100 top networks, along with the network scores, were input into the function *relationshipProb* developed by Milns and collaborators (Milns et al., 2010), which provides an average probability for each arc. These probabilities were then input into their *makeclustersIDhigh* function, which estimates the probability of each arc being part of one out of two categories: low probability or high probability. Each arc was assigned to either a low probability or high probability category in addition to a value corresponding to the uncertainty associated with the classification process (Milns et al., 2010). The high probability category represents those arcs found consistently among the highest scoring networks, while the low probability category represents those arcs found rarely. The uncertainty value is an estimation of the uncertainty associated with the classification of the arcs into either one of the two categories (Milns et al., 2010). The arcs considered as highly probable functional relationships were selected with probability values greater than or equal to 0.5 and an uncertainty value equal or lower than 0.01.

This process still resulted in more variation than desired, thus in order to build a consensus network, the arcs common to 50 repetitions of the above process (starting point of each search, 100 random graphs, then application of the Milns et al. (2010) method to identify highly probable functional relationships) were combined. For each arc common to the 50 repetitions, an average value of the probabilities was calculated and used for building a weighted network. The Markov Blanket of the treatment was identified by applying the *mb* function within the "bnlearn" package. Figure 13 gives the overall view of the steps taken and the decisions made throughout the experimental chapter.

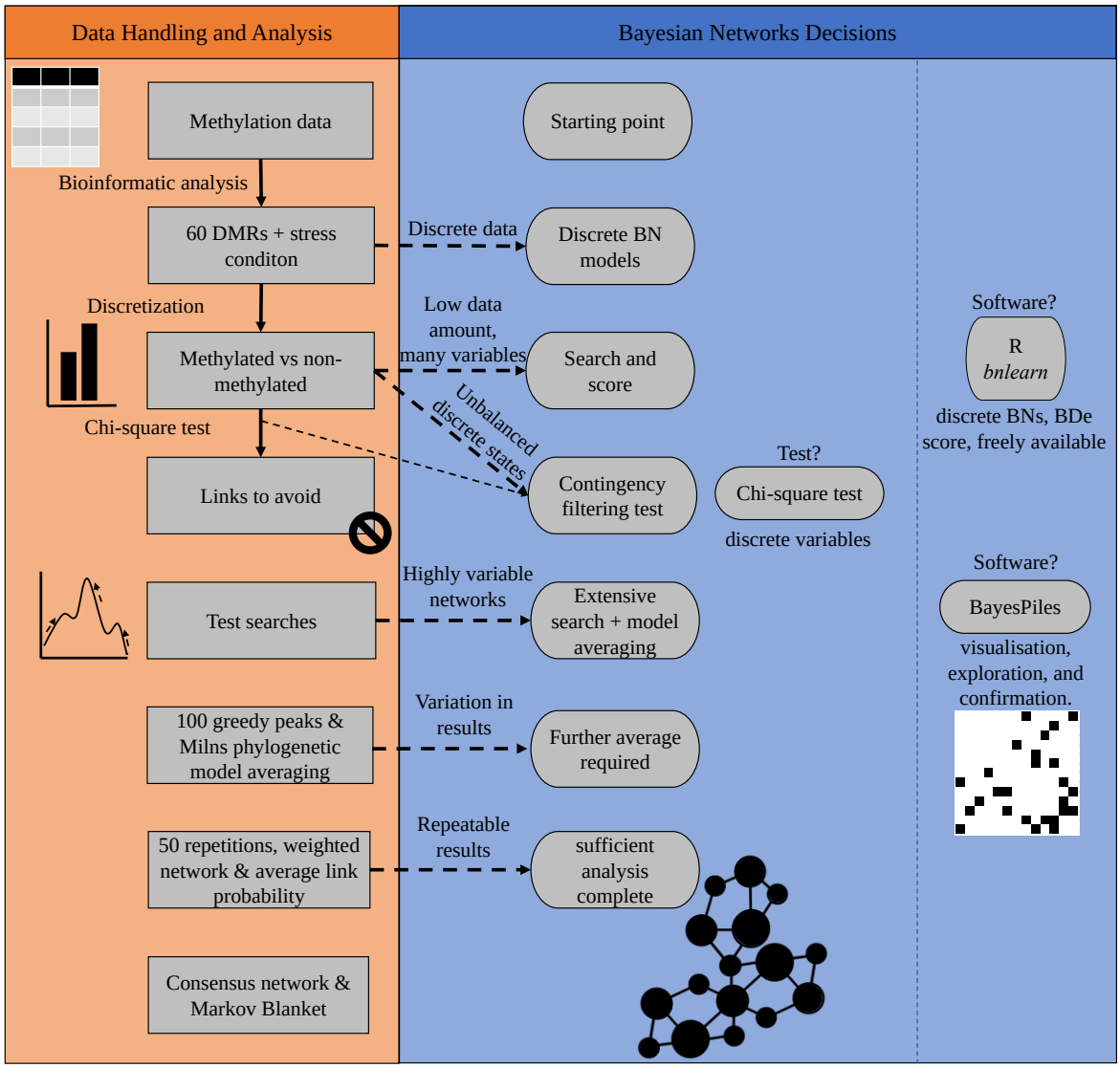

Figure 13. Steps taken and decisions made to build a consensus Bayesian network. The starting point was a dataset consisting of 46 chickens, 22 raised under control conditions and 24 raised under stress conditions. Bioinformatic analysis were performed as described in (Pértille et al., 2017, 2020). Thereafter, a set of 60 differentially methylated regions (DMRs) were selected based on a p-value equal to 0.005. The corresponding methylation values of each DMR were counts (values ranged between 0 and 39). A binary discretization method was implemented, considering that the most frequent value was 0. The software R (and RStudio) was utilised to learn discrete BN. Specifically, the bnlearn package was used, exploring the search space with a score-and-search algorithm and the BDe score. A contingency test (chi-square test) was applied to all possible pairs of variables to create a list of links to avoid, considering that the data had imbalances between the binary states that could lead to the discovery of artefactual links that should not be part of the consensus network. By using the software BayesPiles, it was possible to decide that the search space was complex and building the consensus Bayesian network required a strategic and accurate approach: the combination of a model averaging and the selection of arcs common to all searches into the weighted BN.

6.4. Results.

6.4.1 Bayesian network decisions

A first major choice in BN analysis is whether to use discrete versus continuous models: continuous BNs make use of the numeric value of measured variables but are restricted to additive interactions; discrete BNs use discrete categories for variable values, meaning that numeric data must be "discretised" into ordinal states, but allow for combinatoric interactions (Heckerman et al., 1995). The data consisted of 60 differentially methylated regions (DMRs) measured from 46 male White Leghorn chickens (*Gallus gallus*) plus the experimental condition for each chicken (22 control, 24 stressful condition): these represent the features in the dataset. The experimental condition was a discrete variable. The DMRs were integer values representing the number of sequenced reads for each individual, which represents the methylation level of that specific region per individual; however, the value of 0 (no methylation) was by far the most common, therefore, meaningful discretisation into no-methylation and methylation was a sensible choice. This discrete data combined with the ability of discrete BNs to represent combinatoric interactions, which may be expected in genetic systems (Matthäus et al., 2010), led us to choose discrete BN models.

A BN algorithm works by performing a heuristic search through network structures and selecting structures with high scores under a specific scoring metric. Given discrete BNs, there are a number of scoring metrics to choose among. With the aim of maximising the possibilities of finding novel connections, the Bayesian Dirichlet equivalent (BDe) score was chosen (Heckerman et al., 1995; Yu et al., 2004), which has been shown to be less conservative than others (Bayesian Information Criterion BIC and Mutual Information MI (Yu et al., 2004)). Another choice to make is what software to use to perform the BN analysis, with options ranging from coding it oneself (Affara et al., 2013; Guo et al., 2016) to a variety of free and proprietary platforms (Ciaccio, Wagner, Chuu, Lauffenburger, & Jones, 2010; Kumuthini,

Bessant, Wilson, & Crowther, 2007; Ricard et al., 2019; Shinde et al., 2019; Vasilescu et al., 2017). This choice can be somewhat arbitrary, as the underlying theory remains the same, but will be constrained by one's analysis choices, implementing discrete networks using a BDe score. The R package “bnlearn” was chosen (Scutari, 2011), as free, open-source software which had the desired functionality.

Finally, choices regarding the search process must be made. In order to make informed decisions, iterative exploration of the data and initial search results is required. First, the discrete data was examined, and found that there was an imbalance in discrete states for many of the DMRs (more no-methylation, Figure 14). Because such imbalanced states can create artefactual connections by overrepresented states appearing to be good predictors of each other, regardless of the presence of the rarer states (Milns et al., 2010), the method of contingency test filtering from Milns et al (2010) was applied: a pair-wise chi-square tests was applied, identifying those pairs of variables with chi-square p-values equal to or greater than 0.25 as showing no potential dependence. These were provided to the BN as a list of arcs that must not be considered in the process of building the network (Milns et al., 2010). In total, contingency test filtering identified a total of 960 arcs (of the 3,660 possible arcs) to avoid.

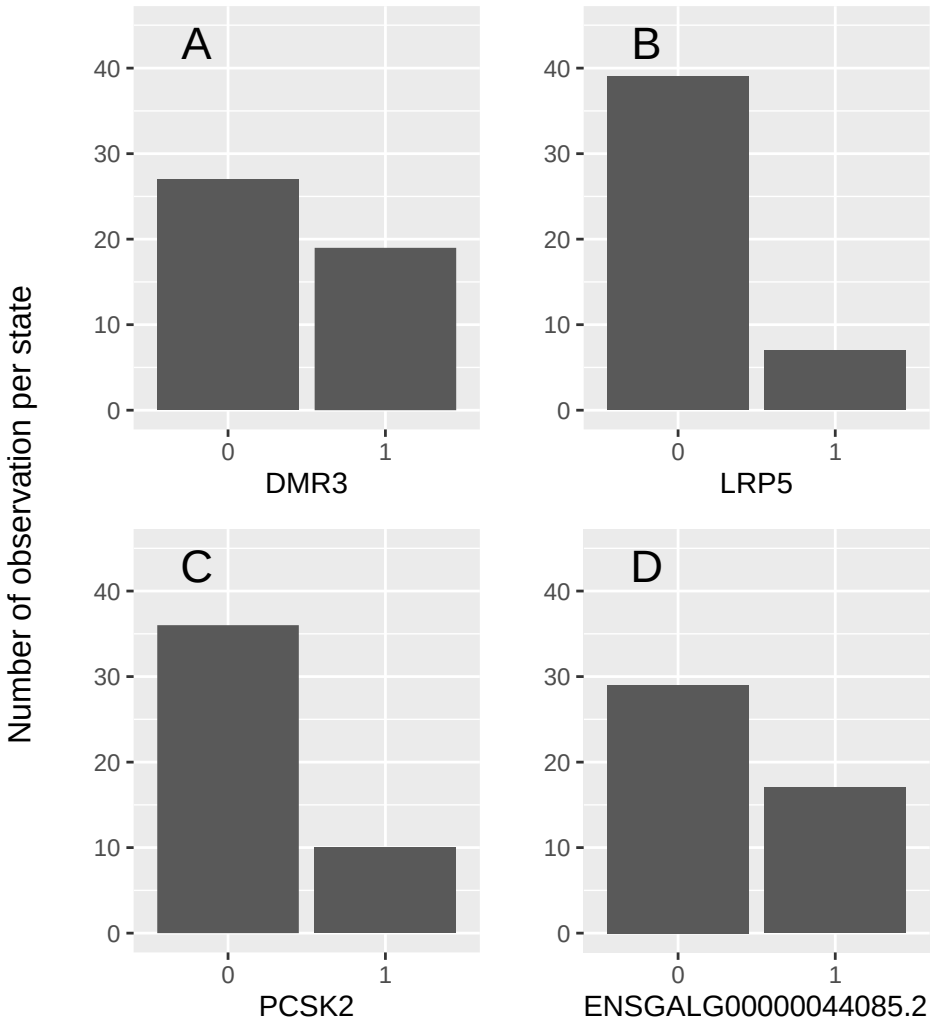

Figure 14. Distribution of four of the differentially methylated regions (DMRs) once a binary discretization method was applied. The state 0 represents values with absence of methylation, the state 1 represents values with presence of methylation. These four DMRs are representative of imbalances between the two states, as zero was the most popular state among different DMRs.

Heuristic searches were initially performed using "bnlearn" R package, finding a large variety in network structure, suggesting that extensive search and model averaging would be the best approach. This was confirmed by the software BayesPiles (Vogogias et al., 2018), which showed highly variable top networks across different searches (Figure 15). Networks similar in score varied strongly in structure. This variation indicates that the top networks found are in different areas of the search space, and not simply fine variations of one general area. Thus, the modelling averaging approach from Milns et al. (2010) was applied, which has been

shown to produce similar sets of highly probability arcs from different collections of top networks (Milns et al., 2010): 100 greedy hill climbs were performed from 100 random starting networks, and applied the Milns model averaging approach to identify highly probable arcs (Milns et al., 2010). As there was still some variation even in these highly probable arcs, this process was repeated 50 times, those arcs common to all searches were selected, and the average probability of the common highly probable arcs across all repetitions was calculated, to produce a final consensus network. Repetition of this analysis showed repeatable results, identifying the same top relationships between DMRs and the same Markov Blanket of the stress condition, thus this was sufficient exploration of the search space.

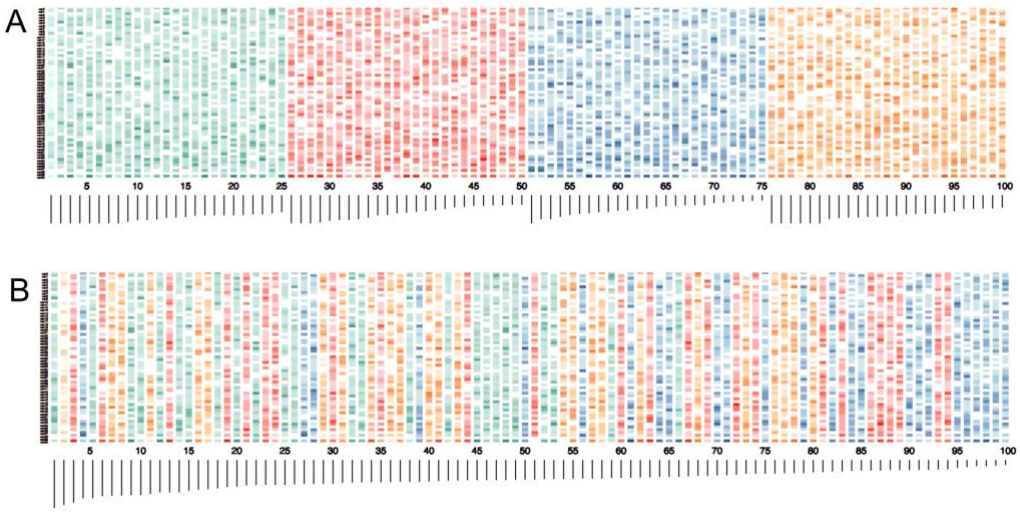

Figure 15. BayesPiles investigation of search space. Top networks found from four separate collections of searches, representing peaks of many different hills in the search space. BayesPiles visualises a summary of network structure as a shaded stack representing out-degree of each node (darker=higher) above a bar representing network score (longer=higher), with networks along the x-axis and nodes along the y-axis. (A) shows the highest 25 networks for four collections of searches (different colours), with highest-scoring network to the left. The strong variation in network structure (different patterns in the shaded bars) indicates that these networks are tops of different peaks in the search space, not the final climb of a single hill. (B) shows the final 25 networks from all four searches combined, sorted by their score. The mixing of colours throughout shows the high variation in search peaks: each collection of searches explored different areas of the search space, finding different high-scoring structures.

6.4.2 Discovered Bayesian network structure

A total of 43 arcs were common to all 50 searches. These arcs and their average probability values of being part of the top 100 networks are shown in Table 4, and the consensus network built with these arcs is shown in Figure 16. The consensus network included 47 out of the 61 features (60 DMRs plus experimental condition). Among these arcs, relationships between DMRs OCLN—DMR7 (distal intergenic region), CANX—TPST2, and FBN1—ENS27231 (unannotated region) had the highest values of probabilities of being part of the consensus network (0.96, 0.86 and 0.83, respectively).

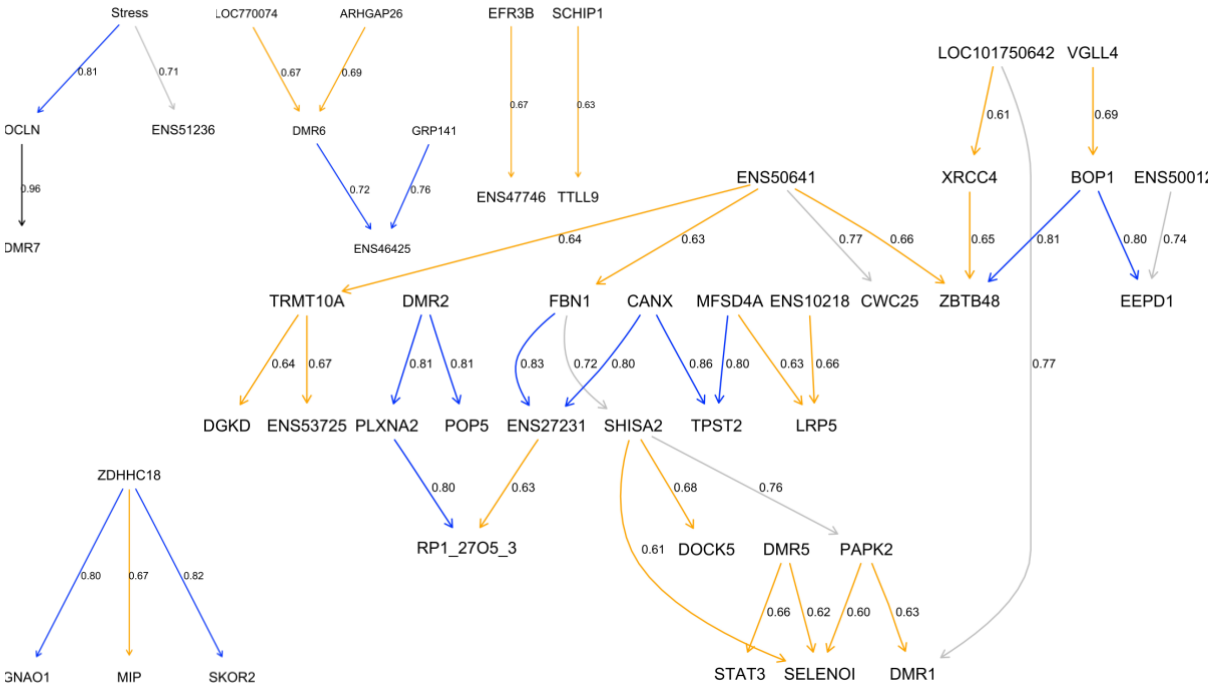

Figure 16. Consensus networks of DMRs. Networks were built with common arcs to 50 searches, each one of these searches consisted of a starting point of 100 random graphs. Features representing the differentially methylated regions (named by related gene or region, see Methods) and the stress conditions are nodes; lines between nodes represent the identified relationships. Note the direction of the arrows do not represent causation, but instead a statistical relationship. Arc labels represent the average probability of belonging to the consensus network, the higher the values, the higher the probability of belonging to a high scoring network. Different colours represent different ranges of probabilities: black: 0.90-1.00, blue: 0.89-0.80; grey: 0.79-0.70; orange: 0.69-0.60.

The application of the Markov Blanket property in order to get the set of parents, children, and spouses in close relationship with the treatment identified only two DMRs, OCLN and ENS51236 (unannotated region), of which the arc between stress and OCLN had the highest average probability value (0.81).

Table 4. Arcs and their corresponding probabilities of being part of a high scoring network. List of arcs identified between differentially methylated regions and with the stress condition, with their corresponding probabilities of being part of a high scoring network. The first column ("arc") is an arbitrary numbering for the arc; the second column ("from") represents the parent node for each arc (arcs from); the third column ("to") represents the child node for each arc (arcs to); the third column ("Average.Probablity") represents the average probability value for each arc of being part of a high scoring network.

	from	to	Average.Probability
1	ARHGAP26	DMR6	0.69
2	BOP1	EEPD1	0.80
3	BOP1	ZBTB48	0.81
4	CANX	ENS27231	0.80
5	CANX	TPST2	0.86
6	DMR2	PLXNA2	0.81
7	DMR2	POP5	0.81
8	DMR5	SELENOI	0.62
9	DMR5	STAT3	0.66
10	DMR6	ENS46425	0.72
11	EFR3B	ENS47746.1	0.67
12	ENS10218	LRP5	0.66
13	ENS27231	RP1_27O5_3	0.63
14	ENS50012	EEPD1	0.74
15	ENS50641	CWC25	0.80
16	ENS50641	FBN1	0.63
17	ENS50641	TRMT10A	0.64
18	ENS50641	ZBTB48	0.66
19	FBN1	ENS27231	0.83
20	FBN1	SHISA2	0.72
21	GRP141	ENS46425	0.76
22	LOC101750642	DMR1	0.77
23	LOC101750642	XRCC4	0.61
24	LOC770074	DMR6	0.67
25	MFSD4A	LRP5	0.63
26	MFSD4A	TPST2	0.80
27	OCLN	DMR7	0.95

28	PAPK2	DMR1	0.63
29	PAPK2	SELENOI	0.60
30	PLXNA2	RP1_27O5_3	0.80
31	SCHIP1	TTLL9	0.63
32	SHISA2	DOCK5	0.68
33	SHISA2	PAPK2	0.76
34	SHISA2	SELENOI	0.61
35	Treatment	ENS51236	0.71
36	Treatment	OCLN	0.81
37	TRMT10A	DGKD	0.64
38	TRMT10A	ENS53725	0.67
39	VGLL4	BOP1	0.69
40	XRCC4	ZBTB48	0.65
41	ZDHHC18	GNAO1	0.80
42	ZDHHC18	MIP	0.67
43	ZDHHC18	SKOR2	0.82

6.5. Partial discussions.

Behind biological systems lies a series of complex and intricate relationships among features (Guzmán et al., 2017; Milns et al., 2010). The application of BNs can be a useful approach to discover, identify, and unravel hidden patterns within the data, and gain insights into a biological area of knowledge (Felipe et al., 2014; Needham et al., 2007). However, there is little practical guidance for how to make choices among the array of possibilities within a BN analysis. A practical application of BNs was undertaken to answer a particular question in poultry epigenetics, while clearly stating the analysis choices. The reasoning behind using a discrete, rather than a continuous, BN due to the distribution of the data, how the BDe score was chosen, and the software applied were explained. The analysis of the dataset's discrete states and the choice to use chi-square contingency-test filtering to avoid artefacts from imbalanced discrete states were also explained. The exploration of the search space structure for the question was shown, including using the specialised software BayesPiles (Milns et al., 2010), which revealed the space to be highly varied and thus to require complex model averaging techniques developed in a similarly varied search space were applied (Milns et al., 2010), and added further refinements of combining multiple searches. The clarity surrounding the current choices might be helpful for others beginning a BN analysis as a sort of roadmap.

The Markov Blanket of the stress condition together with those DMRs showing the highest probability values of being part of the consensus network appear related to the functional structure of the brain and a possible link with the immune system. Starting with OCLN, Occludin, this gene showed the highest probability value and it belonged to the Markov Blanket of the stress condition. OCLN is a gene whose major functionality is associated with intracellular tight junctions and adhesion, defining a selective barrier and limiting the exchange of substances and/or cells in different tissues such as the chicken ovary, the chicken intestinal mucosa, or the human brain (Du et al., 2017; Furuse & Tsukita, 2006; Schuster, Schmierer,

Shkumatava, & Kuchler, 2004; Stephens & Johnson, 2017; Zhang et al., 2012). In the chicken ovarian follicles, OCLN plays a role in allowing or preventing the exchange of yolk material, especially during the first stages of the formation of the follicles, considering that the expression values were increased (Schuster et al., 2004; Stephens & Johnson, 2017). In case of the brain, OCLN and other genes are involved in the permeability of the blood brain barrier, as its integrity is crucial for the correct functioning of the central nervous system (Du et al., 2017; Furuse & Tsukita, 2006). In human patients suffering from a fatal heat stroke, increased expression values of OCLN were found, and authors suggested that it could be aimed at restoring junctional complexes and the barrier function as a compensatory mechanism (Du et al., 2017). Considering that the stress response is initially triggered in the central nervous system, it is possible that OCLN is playing a key role protecting the integrity of the blood brain barrier to prevent any nervous disfunction, that would be crucial when dealing with the influence of a stressor.

The arc between CANX and TPST 2 was among the arcs with the highest probability values. The biological functionality of CANX, Calnexin, can be divided into two major categories as it is linked to the immune system as well as to the blood brain barrier (Jung et al., 2018; Lai, Teodoro, & Volchuk, 2007; Sekelova et al., 2017; Sperandio, Gleissner, & Ley, 2009). Chickens inoculated with *Salmonella* Enteritidis as an immune challenge increased the abundance of CANX in heterophils (a subpopulation of leukocytes) (Sekelova et al., 2017). Together with other proteins, CANX belongs to the endoplasmic reticulum proteins and their functionality comes into play when the unfolded or misfolded proteins exceed the capacity of chaperones or when the luminal conditions are not optimal for the correct processing of new proteins (Lai et al., 2007). Regarding CANX functionality in the brain, Jung et al (2018) found that this gene plays a major role in multiple sclerosis and its equivalent in mice, as the loss of CANX increased the resistance of the blood brain barrier, avoiding the infiltration of cells

belonging to the immune system and the induction of inflammation markers (Jung et al., 2018). The other DMR interacting with CANX was TPST 2, a tyrosyl protein sulfotransferase that, and together with TPST 1, are in charge of the correct functioning of P-selectin glycoprotein ligand-1 (PSLG-1) by transferring tyrosine residues (Ouyang & Moore, 1998; Sperandio et al., 2009; Yu, Hoffhines, Moore, & Leary, 2007). PSLG-1 is expressed on leukocytes and promotes binding and adhesive interaction with other selectins that may lead to inflammatory disorders as a consequence of a potential pathological recruitment of leukocytes (McEver & Cummings, 1997).

It is important to consider that the DMRs were evaluated in the red blood cells of chickens raised under either control or stress conditions. Throughout the previous experimental chapter, it was mentioned that the exposure of chickens to stress can affect the expression pattern of genes in different tissues. For example, some HSP were identified by BN in the brain (chapter 4) whereas in the literature they were found in other tissues such as the spleen, heart, or muscle (Guo et al., 2020; Xie et al., 2014). Particularly, HSPH1 was found in the brain (experimental chapter 4) and in the spleen of chickens (Guo et al., 2020). Considering that blood is systemically circulating through the organism, it is difficult to tell whether the differential methylation patterns start occurring in the red blood cells themselves or whether these changes are reflecting those occurring in specific tissues (Pértille et al., 2017, 2020). This is another example of an application of BNs in knowledge discovery, allowing the possibility to define new hypothesis to be tested in future experiments.

Among the Markov Blanket as well as the arcs with the highest weight values there were 3 DRMs whose function and/or annotation is still yet unknown (DMR7, ENS51236, and ENS27231). Finding highlights two different advantages of implementing BNs: on the one hand, studies focusing only on bioinformatic analysis would generally ignore these DMRs or genes, because the functionality of them will not be found in sources such as KEGG pathways

or GO terms. On the other hand, the power of BN algorithm discovered novel markers that might be worth exploring. Learning the structure of a BN with a set of highly significant genetic features can be the starting point of future research. Instead of focusing on the bigger picture that bioinformatic studies provide, analysis of only a reduced number of features would be more accurate to gain a further insight into the stress phenomenon.

The stressful condition, in this particular study, was directly connected to only one epigenetic feature, OCLN, while the Markov Blanket consisted of two epigenetic features. It is then plausible to ask whether these two DMRs can be explored as biomarkers of stress in chickens. Considering these finding as the starting point, future studies can be aimed at evaluating the expression and/or methylation patterns of only these two genetic features under two experimental conditions, non-stress and stress. Thereafter, knowledge can be transferred into other fields such as animal welfare and poultry production. For example, one of the main principles of animal welfare is the absence of distress in association with a comfortable environment (Ducatelle et al., 2018; Fallahsharoudi et al., 2017; Webster, 2016). Stress can be highlighted as one of the major problems faced by the poultry industry nowadays, and the knowledge discovered by BNs can be further used to develop breeding protocols and genetic lines (Lara & Rostagno, 2013; Renaudeau et al., 2012). Even though in this particular study the condition was stress, it is important to mention that the condition could be of any other nature, such as gender, male vs female; phenotypes, ancestral vs domesticated chickens; or even different stages in life, juvenile vs adult (Bélteky et al., 2018, 2016; Elfwing et al., 2015). In this context, the approach implemented in this study can be applied in genetics and epigenetics as a first approximation to gain basic knowledge in regard to a particular condition, with potential implications in applied science.

7. Discussion.

Throughout this thesis, a general overview of poultry and genetics has been explored. Two main genetic features and their relationship with the stress phenomenon have been studied: differential expression patterns and differentially methylated regions. Genes and methylated regions of the DNA have been identified as key factors relevant to the stress phenomenon, revealing hidden relationships among them as well as with between them and the stress condition. From a general and broad perspective, throughout the thesis there have been several challenges to be faced, considering the complexity of the stress phenomenon as well as the fact of working with biological big data. These challenges came from the collection of data, the pre-processing, and the process of learning the structure of the BNs. In the following sections, these challenges such as data collection, discretization of the data, strategies and approaches to BNs will be addressed, followed by the genetic and biological implications of the findings, and finally, the possible limitations and future research will be discussed to close this chapter as well as this thesis.

7.1. Chicken vs human-related studies.

BNs are mathematical tools that can be applied in a variety of research fields, from molecular biology, cancer research, psychology and psychopathology, to ecology and environmental sciences (Agrahari et al., 2018; Balov, 2013; Blanchard, Roskam, Mikolajczak, & Heeren, 2021; Faverial et al., 2016; Hartnack et al., 2019; Meier et al., 2020; Milns et al., 2010; Mitchell et al., 2021; Peters et al., 2021). In genetics and epigenetics, independently of the animal model (chickens, rodents, humans), BNs have also been applied with the aim of associating phenotypes (healthy vs unhealthy, or control vs stress) and a given set of genes (Agrahari et al., 2018; Balov, 2013; Djebbari & Quackenbush, 2008; Li et al., 2010). In this particular field of research, latest advances in technology have allowed the possibility to measure hundreds of thousands genetic features as well as reducing the economic costs of such measurements, increasing the availability of these technologies to measure the effects of

several conditions or treatments. In the era of big data, massive amounts of biological data are being collected and deposited in publicly available repositories, such as Gene Expression Omnibus (GEO), Array Express, The Cancer Prevention and Outcomes Data (C-POD) Shared Resource, the English Lung Cancer Database (LUCADA), among many others. These online repositories are there, on the one hand, so that researchers can submit their data coming from different studies; and on the other hand, so as other researchers can reuse and re-analyse the data, and, as a consequence, they can contribute with new findings in regards to a particular research topic. Even though studies in a particular field (e.g., cancer or stress) share common characteristics, such as the technologies used to measure genetic/epigenetic features (e.g., microarray or RNA-sequencing), there are some clear differences between poultry species and other models, such as rodents or humans.

One of the main differences between animal models is related to experimental designs. Studies carried out in cancer research have simple experimental designs, generally focusing on one particular type of cancer and the identification of differences in the patterns of patients classified according to their health status (with or without cancer) (Balov, 2013; Koussounadis, Langdon, Harrison, & Smith, 2014; Xia et al., 2019). Even though some studies carried out in chickens might focus on one experimental factor, such as stress (Pértille et al., 2020), other studies might have complex experimental designs, evaluating more than one factor with two possible levels, stress and domestication, or stress and life stages (Bélteky, Agnvall, & Jensen, 2017; Ericsson et al., 2016; Løtvedt et al., 2017). In addition to the complexity of the experimental designs, researchers have focused on the impacts of different environmental conditions on different tissues, such as spleen, liver, brain, among others (Bélteky et al., 2016; Li et al., 2010; Saelao et al., 2018; Van Goor et al., 2017). As a consequence, and as another main difference between chicken and human studies, the number of individuals is clearly affected by the experimental designs, as a small number of chickens can be associated with a

particular combination of environmental conditions. In this sense, it is possible to identify a set of genetic features when looking for differences in the expression patterns between two conditions using a reduced number of individuals, especially in poultry science. However, in this particular thesis, the first step was to collect data, and using a small number of individuals associated to a specific condition coming from only one study might have an impact on the search space when leaning the overall structure of the BN. Therefore, collecting and reusing publicly available databases was the strategy implemented to overcome this challenge, combining datasets coming from different studies and experimental designs. However, and due to the previously mentioned differences between animal models, the number of individuals can be considered relatively small compared to studies evaluating, for example, cancer in humans (Balov, 2013; Berkan Sesen, Nicholson, Banares-Alcantara, Kadir, & Brady, 2013; Cruz-Ramírez, Acosta-Mesa, Carrillo-Calvet, Nava-Fernández, & Barrientos-Martínez, 2007). This fact did not prevent us from learning the structure of the BNs, but it did have an impact on the search space, requiring extra steps to build the consensus network.

7.2. Dealing with a challenging field of research.

Once the datasets were place, two main challenges were faced depending on the type of data (gene expression or methylated regions). The first challenge encountered was the need for variables with discrete states, reducing the number of states before learning the structure of the networks. BN algorithms can handle continuous data, nonetheless, this possibility is limited to additive interactions, which are not necessarily what is always expected in genetic systems. Additionally, the process of dealing with complex non-linear relationships together with the statistical and computational power might be difficult to achieve as well as it could be a time-consuming process (Heckerman et al., 1995; Milns et al., 2010; Yu et al., 2004). Even though working with continuous data might result in further details into the relationships of a given set of variables as well as keeping the data with their original values, discretizing the data has

the advantage of reducing the number of states, consequently increasing the statistical and computational power as well as reducing and filtering noise (Balov, 2013; Heckerman et al., 1995; Milns et al., 2010; Yu et al., 2004). This last advantage was useful, considering that the data came from different experiments, thus the noise associated to each particular study could be filtered and reduced. In terms of the discretization method, when working with BNs it is desirable to work with variables which categories are evenly distributed, especially due to the statistical power and to avoid imbalances between the categories (Heckerman et al., 1995; Milns et al., 2010; Smith, 2010; Yu et al., 2004). Having imbalances between categories will lead to the discovery of links due to an artefact of the algorithm instead of representing a true link. In this thesis, both gene expression datasets were discretized into evenly distributed categories with a quantile discretization method. However, when working with the epigenetic dataset, it was not possible to work with this type of method due to the distribution of the data. This dataset consisted of counts, representing the number of differentially methylated regions that a chicken had driven by the social isolation stressor. Values ranged from 0 to 39, but the most frequent value was zero among the DMRs, having an initial imbalance between the discrete states. Consequently, dividing the data into evenly distributed categories was not feasible. The strategy used in this dataset was a binary approach: considering that a value of zero meant the absence of methylation, while the rest of the values meant the presence of methylation. The implementation of the binary discretization method partially overcame the initial challenge with unevenly distributed variables.

The epigenetic dataset also represented a further challenge, especially due to the distribution of the original values. Dealing with unevenly distributed variables might affect the overall process of learning the structure of the network, requiring an extra step to filter arcs between nodes that showed no possible dependence. Previous studies in ecology and psychology have implemented filtering methods to remove arcs that should not be considered

when learning the structure of the network (Meier et al., 2020; Milns et al., 2010; Mitchell et al., 2021). This is because a relationship might be identified between two variables based on their most popular state, resulting in an artefact of the algorithm rather than a true interaction. In order to overcome this challenge, a previously published contingency test was applied so as to identify a set of arcs that must not be included in the network as there is no evidence of dependence between two given nodes (Milns et al., 2010; Mitchell et al., 2021). Considering that the variables had discrete states, a chi-square test was applied to all possible combinations of the given 60 DMRs and the stressful condition. As previously implemented by Milns et al. (2010) and Mitchell et al. (2021) a set of arcs was identified considering a threshold of 0.25, banning these connections when learning the network. The epigenetic dataset was the only one requiring the inclusion of prior information in the form of a set of arcs to be blocked.

As discussed in the previous section, the number of observations played an important role when learning the structure of the genetic and epigenetic BNs. Experimental designs of previously published studies had a major impact on the amount of data available to build the datasets used in this thesis. Additionally, the number of variables also had an impact on the search space, considering that the higher the number of variables, the more complex and computationally expensive is the search for the network that best fits the data (Chickering, 1996; Cooper, 1990; Vogogias et al., 2018). Solving this challenge required not only exploring fields outside genetics, bioinformatics, and machine learning, but also creativity together with scientific accuracy. Throughout this thesis, different strategies have been implemented with the aim of finding the consensus network. In terms of BNs, the simplest of these strategies is to consider the consensus network as the top highest scoring network, if the algorithm finds the same structure every time the space is searched (Vogogias et al., 2018). In terms of complexity, the next strategy would be to combine the results of the n highest scoring networks from n different searches; n being a discrete number such as 10 or 100. This can be considered

as the consensus BN, as it involves the combination of multiple searches into one network (Vogogias et al., 2018). If the structure of the network is slightly different between searches, the following strategy would be to combine n top highest scoring networks or consensus networks into weighted networks. This approach requires the identification of arcs present (or absent) across different searches and then setting a threshold to filter arcs found in few searches, implemented as a strategy to get rid of arcs that might have been discovered as an artefact of the heuristic random search (Peters et al., 2021; Rodgers et al., 2019; Vogogias et al., 2018). In this thesis, weighted networks were used in the genetic datasets, combining 50 consensus BNs in the case of the brain, stress, and microarray dataset, and 100 consensus BNs in the case of the spleen, stress, and RNA-sequencing dataset. In both cases, the threshold was set to consider those arcs present in at least 50% of the searches.

The most complex search space was the one corresponding to the epigenetic dataset, that required the combination of several strategies. The starting point was 100 BNs. Thereafter, a matrix was created considering the presence or the absence of a particular arc across the different networks. The score of each particular network was also included as part of this matrix, as it was needed to implement an average method previously published by Milns and colleagues (2010). This method calculated the average probability of each arc of being part of the top 100 networks by using the matrix of presence/absence of arcs as well as the score based on a phylogenetic tree approach. This approach treats the presence of arcs in a network as features to compare the similarities between networks as in phylogenetic trees, followed by a regression on the scores of the networks controlled by the correlation patterns of the phylogenetic tree (Milns et al., 2010). The aim of this approach is to identify an average probability of each arc being in a high-scoring network. These probabilities are clustered using a Gaussian mixed model to identify highly probable arcs (Milns et al., 2010). Finally, and in order to identify highly probable functional relationships, the probability of each arc being part

of one out of two categories was calculated: each arc was assigned to either a lowly or a highly probable category in addition to a value corresponding to the uncertainty associated with the classification process (Milns et al., 2010). As challenging as this particular search space was, a further and final step was needed: building the consensus network by combining the results of 50 searches and using those arcs present in at least 50% of the searches.

Based on the strategies and approaches implemented throughout this thesis, not only the number of observations play an important role in regards to the complexity of the search space, but also the number of variables included when learning the overall structure of the network (Chickering, 1996; Cooper, 1990; Vogogias et al., 2018). Therefore, it is important to consider these two factors prior to learning the structure of the network, as they would be influencing the complexity of the search space. A complex search space shaped by the number of observations and variables would require a series of concatenated steps in order to build the consensus network.

7.3. A bigger picture of the biology behind the results.

Throughout the thesis, different genetic and epigenetic features have been identified to be associated with stress. Even though the nature of this thesis is mostly exploratory, our results might be understood in the context of poultry management as well as in domestication and artificial selection.

When evaluating the effect of stress on the physiology of avian species, there are several indicators of stress. The most common is the measurement of corticosterone, a glucocorticoid which plasmatic concentration increases as a consequence of the activation of Hypothalamus-Pituitary-Adrenal axis of the neuroendocrine interplay (Cantet et al., 2021, De Kloet, 2003; Dhabhar & McEwen, 1997; Dohms & Metz, 1991; McEwen et al., 1997; Selye, 1950, Wickramasuriya et al., 2022). Considering that the activation of the neuroendocrine system has a direct impact on the immune system, an alternative indicator of stress has been explored: the

Heterophil to Lymphocyte ratio (Mashaly et al., 2004; Nazar, Estevez, Correa, & Marin, 2017; Scanes, 2016). The correlation of these two indicators is based on the fact that glucocorticoids promote the migration and mobilisation of immune cells from tissues and organs to the blood stream and vice versa. As a consequence, a bird exposed to stress will show increased counts of heterophil while decreased counts lymphocyte in blood smears, with an overall increase in the ratio between Heterophil/Lymphocyte (Huff, Huff, & Balog, 2005; Scanes, 2016). Even though these two indicators are mostly used as biomarkers of stress, several other side effects of stress have been reported involving changes in the gut microbiota, the health of the intestine, alteration of immune variables, among others (Burkholder et al., 2008; Nazar et al., 2015; Nazar et al., 2017; Quinteiro-Filho et al., 2012, 2010; Shi et al., 2019).

Exploring genetic or epigenetic biomarkers in poultry species as indicators of the stress phenomenon is not as advanced as in other animal models, especially rodents or humans. So far, studies have focused on the identification of differentially expressed genes or differentially methylated regions driven by a particular condition (Fallahsharoudi et al., 2017; Goerlich et al., 2012; Guo et al., 2020; Pértille et al., 2017, 2020). However, the set of genetic or epigenetic features can be quite extensive and trying to identify their corresponding biological pathways or gene ontology terms can be a complex task. This is because of the volume of genes having differential expression or methylation patterns, including those that might not have a gene symbol or other genetic attribute that can help tracing the genes in other databases. Additionally, understanding the biology behind such an extensive list of genes or DNA regions can be challenging, as the pathways could be either general responses related to the immune system, metabolism, neurogenesis, or specific responses such as cell signalling and cascades or particular mechanisms of the immune system (Guo et al., 2020; Pértille et al., 2017).

Although there are some differences between animal models, such as experimental designs and the number of animals tested previously discussed, some other differences can be

related to the current scientific community, potentially influencing this gap between models. Even though ethical approvals must be in place before carrying out experiments in any research involving animals or humans, medical research can often take advantage of samples taken in the course of medical treatment, e.g., cancer biopsies, with generic permission given for storage in tissue banks and thus providing a base of samples when a new study is designed. Additionally, more data repositories are publicly available, such as the Cancer Prevention and Outcomes Data (Wisconsin database) or the English Lung Cancer Database (LUCADA), and they are daily updated with new entries. On the other hand, experiments carried out in poultry species are quite specific compared to the ones in humans, as a small number of chickens are being studied under specific stressful conditions, such as heat stress or social isolation stress, on specific tissues, such as brain, breast, or spleen, and often including extra experimental conditions such as domestication or different generations. Moreover, evaluating the same condition and/or experimental designs cannot be considered by researchers, as the overall idea has already been published and the findings will not be original. The lack of originality represents a problem, especially when dealing with funding applications or when communicating the results to the scientific community, considering that experimental designs already evaluated do not contribute to knowledge discovery. Poultry scientists, unlike medical researchers, usually need to think and develop new experimental designs as well as novel ways to evaluate the same phenomenon, stress in the case of this particular thesis.

If we were to compare medical research with poultry science, medical research has identified some useful genetic biomarkers associated with specific types of cancer as a way to diagnose and/or prevent patients from suffering advanced stages of the disease. Some examples can be the mutation of the BRCA1 gene, that has been used a biomarker of hereditary breast cancer as well as breast-ovarian cancer, while the prostate-specific antigen, an enzyme produced by the prostate, has been used as a biomarker of prostate cancer (Carser et al., 2011;

Hernández & Thompson, 2004; Rosen, Fan, Pestell, & Goldberg, 2003; Tkac et al., 2019). In this sense, it is plausible to think that stress, as cancer, is a broad and general phenomenon with common mechanisms to different types of stressors. In this thesis, genes and epigenetic features related to oxidative stress, apoptosis, and tight junctions have been identified and they could be used as hallmarks of stress. However, more studies with an in-depth approach to each specific type of stressor and its impacts on specific tissues are required.

On a bigger scale, both in terms of implications and time, it is worth reflecting on domestication, breeding programmes, and artificial selection. Poultry species have undergone an extensive domestication process for more than 8000 years, and they are currently undergoing artificial selection programmes. In ancient societies, domesticated birds were initially "selected" based on two interrelated characters: their tameness and their ability to survive in this new environment close to humans (Agnvall, Jöngren, Strandberg, & Jensen, 2012; Ericsson et al., 2016; Mignon-Grasteau et al., 2005). Thereafter, and especially during the last decades, poultry species are undergoing breeding programs and artificial selection with the aim of increasing production and performance, with underlying economic purposes. Focusing on chickens, current breeds can, for example, produce 5 to 7 eggs a week or achieve slaughter weight at around 42 days of age (Cheng, 2010; Jackson & Diamond, 1996; Ma et al., 2018). However, selecting those birds towards economic-productivity goals might have led poultry species to become more susceptible to certain conditions that do not meet the conditions outside those required for optimal productivity. It is in this scenario that stress comes into play in several forms such as transportation from the hatching facilities to the farms, temperatures above or below thermoneutral, social interactions with conspecifics, exposure to pathogens among many others (Burkholder et al., 2008; Guzmán et al., 2017; Kuenzel & Jurkevich, 2010; Lay et al., 2011; Nazar, Magnoli, Dalcero, & Marin, 2012; Rosales, 1994).

Studies have been done in the recent years with the aim of solving the current problems that the poultry sector is facing. Some examples of breeds selected based on specific characteristics could be the slower growing broilers, chickens with low fear of humans, and the stress and/or disease resistant chicken breeds (Bélteky et al., 2018; Dixon, 2020; Katajamaa & Jensen, 2020; Park et al., 2019; Van Goor et al., 2017). Dixon (2020) carried out a set of experiments to evaluate the performance as well as some welfare indicators in two types of breeds: commercial faster growing vs slower growing breeds (Dixon, 2020). Results showed that the slower growing breed required two extra weeks to reach slaughter weight; however, they showed an improved overall welfare status considering some behavioural and health indicators. For instance, the slower growing breed performed a wider range of behaviours, rarely seen on the faster growing breeds, such as perching (Dixon, 2020). Additionally, the slower growing breed had better scores on the feather coverage, breast feather cleanliness, and hock lesions indicators (Dixon, 2020). Selecting birds based on their low or high fear to humans could be considered as the first step of the domestication of poultry species (Bélteky et al., 2018; Katajamaa & Jensen, 2020). Studies on the relative brain size in comparison to the body weight and an epigenetic features showed differences between chickens with low fear and high fear to humans (Bélteky et al., 2018; Katajamaa & Jensen, 2020). Katajamaa and Jensen (2020) also reported that chickens with low fear to humans coped better with a fear habituation test, showing lower scores than their high fear to human counterpart. As one final example, the Fayoumi chickens can be mentioned: this breed is indigenous to Egypt and it has undergone a natural selection process due to the harsh environmental conditions (Lamon et al., 2014). The Fayoumi chickens have been used as a stress-resistant and disease-resistant breed in studies that compared this breed with Leghorns chickens (Deist et al., 2017; Wang, Lupiani, Reddy, Lamont, & Zhou, 2014). Both breeds were exposed to high environmental temperatures (Van Goor et al., 2017; Wang et al., 2018) or to an immune challenge such as Avian Influenza virus

or Newcastle disease virus (Deist et al., 2017; Wang et al., 2014). These studies showed that these two breeds have physiological differences as well as differential expression patterns depending on the challenging condition.

Even though these specific characters have been selected based on the phenotype of poultry species, it is also possible to artificially select characters based on the genotype. In addition to microarray technologies and RNA-sequencing, the study of single nucleotide polymorphisms (SNP) and quantitative trait loci (QTL) could be of interest to identify regions within the genome that can increase the resistance or resilience of chickens to stress (Fallahsharoudi et al., 2017; Ma et al., 2018; Renaudeau et al., 2012). The complexity of this approach is higher compared to the artificial selection based on a phenotypic trait, as it would initially require a deep understanding of the stress phenomenon and the genetics behind it. Additionally, the crossing between individuals should be focused on increasing the overrepresentation of the genetic features of interest. However, it is target-specific considering that the artificial selection is oriented towards a particular genetic trait (Cheng, 2010; Renaudeau et al., 2012).

7.4. Future directions in the genomic field.

Working with genetics involves complexity and accuracy, different approaches, and possibilities, as well as a variety of genetic features to work with. Genetic research can be based on single nucleotide polymorphisms (SNPs), differentially expressed genes, quantitate trait loci (QTL), differentially methylation patterns, among others (Aslam et al., 2012; Ericsson et al., 2016; Fallahsharoudi et al., 2017; Li et al., 2013). Some of these genetic traits can be studied as individual effects such as the single modification of the base of a gene in the case of SNPs or the expression values of a particular gene. The exploration of these individual genetic features associated to a particular condition of interest can be done with bioinformatic analysis such as t-tests or genome-wide association analysis. The study of these individual genetic

features has identified key SNPs or genes that can be utilised in estimating the effects of these features on a particular phenotype or predict the outcome based on individual values (Goddard, Kemper, MacLeod, Chamberlain, & Hayes, 2016). Considering some of the outcomes of this thesis, some of the identified genes or DNA regions can be used with this aim, classifying chickens under control or stress condition when the breeding conditions are unknown or using these genes or DNA regions as biomarkers of stress, with the possibility of implementing them in the management of decision-making.

However, genetics is more complicated than the effects of individual genetic features, and SNPs, genes, or DNA regions can potentially interact with each other, requiring further exploration, considering epistatic effects or haplotype-based genome-wide associations (Aslam et al., 2012; Goddard et al., 2016; Li et al., 2013; Zhang et al., 2017, 2020). Epistasis is the branch of genetics in charge of studying the interactions between SNPs, genes, or QTLs and it is based on the fact that the individual contribution of genetic features might not have the same outcome as their interactions (Cordell, 2002; Howard et al., 2017; Li et al., 2013). In this scenario, Fangge et al (2013) found that SNP-SNP interactions were contributing to the abdominal fat weight of chickens, where individual SNPs might interact with up to 7 other SNPs (Li et al., 2013). 50 pathways were related to these SNPs, some of them associated with obesity, fat droplet formation, insulin-signalling pathways, among others. Zhang et al. (2017) studied the effects of the interactions between SNPs on testis growth and reproduction (Zhang et al., 2017). The findings revealed that over 400 interactions between SNPs were occurring, and one SNPs can interact with up to 8 other SNPs. The biological functionality of these genes pointed towards testis development, motility and viability of sperm, sperm maturation and fertilization, spermatogenesis, among others (Zhang et al., 2017).

Genetic features might be displaying interactions and therefore contributing to a phenotype as a whole rather than as an individual feature, highlighting the importance and the

power of the application of BN approaches to genetics and bioinformatics. On the one hand, BNs can identify individual genetic features associated with a particular condition, in the case of this thesis, stress. On the other hand, BNs can also display relationships and interactions between genetic features that can be related to epistatic effects. Applying BN approaches allow the possibility not only to discover genes that can be utilised to classify or predict outcomes depending on the condition of interest but also define the starting point for future studies with the aim of exploring the actual effects of the discovered interactions on the phenotype.

7.5. Limitations and applications outside genetics.

Even though BNs have many advantages as expert systems that can be used to predict and/or classify the outcome of new data based on previous knowledge and/or data available, or to discover interesting and informative interactions and patterns within a given set of variables, they do have limitations and drawbacks. To start with, the quality and the quantity of the data have been identified as a one major issue when working with BNs. Lupo et al (2013) and Hartnack et al (2019) highlighted the importance of having accurate measures, avoid missing values, and the size of the dataset, as they could introduce noise or potential artefacts when learning the structure of the network (Hartnack et al., 2019; Lupo et al., 2013). Another drawback highlighted by several authors (Comin et al., 2019; Hartnack et al., 2019; Hidano et al., 2015) is related to the importance of the discretization of the data and the way they are encoded. There are several discretization methods that have different strategies, dividing the data into categories with different or equal number of observations per category, depending on whether there is a numeric threshold or not (interval vs quantile method, respectively). Additionally, multiclass variables can also be encoded as binary variables, creating dummy variables for each class of the original variable. With this particular strategy, the number of variables will consequently increase, requiring not only computer power but also statistical power, requiring a relatively large amount of data (Comin et al., 2019). Even though the

discretization of the data is generally implemented to deal with such complex non-linear relationships between a given set of variables, Parsons et al (2005) highlighted that BNs outcomes might be affected, leading to inconsistency and a diminished capability of explaining complexity (Parsons et al., 2005).

Application of BNs in future studies can be developed in a wide range of poultry topics considering studies implemented in fields such as psychology, ecology, molecular biology, or the combination of multidisciplinary approaches (Balov, 2013; Blanchard et al., 2021; Milns et al., 2010; Mitchell et al., 2021; Peters et al., 2021). All in all, a multitude of possibilities can arise to apply a BN approach in poultry research. Some of them have been mentioned throughout the thesis, such as bacteriology and epidemiology, with the aim of understanding the complex relationships between microorganisms, genes, and antimicrobial resistance (Hartnack et al., 2019; Hidano et al., 2015). A field not yet explored regarding poultry research and BNs is behaviour. In humans, several aspects of psychology and psychopathology have been explored such as eating disorders, sports and emotions, or parental burnout (Blanchard et al., 2021; Meier et al., 2020; Peters et al., 2021). Poultry species are currently being studied in different aspects of their behaviour, such as their cognition, locomotion, or their fear response (Guzmán et al., 2017; Hedlund et al., 2021; Jones et al., 1992; Lábaque, Kembro, Luna, & Marin, 2013; Zidar et al., 2017). Application of BN approaches in this field might unravel behavioural patterns varied in nature, displaying relationships and interactions between the outcomes of particular behavioural tests, but all in all, they will provide further insights into how poultry species behave in commercial settings, possibly recognising strategies to improve animal welfare.

Another possible field of interest would be related to productive phenotypes and an overall view of the physiology of different systems within the organism. A holistic approach can be explored with the aim of studying how several systems interact and relate with each

other, considering for example the immune-neuroendocrine interplay. Indicators coming from different systems, such as expression values of genes, plasmatic concentrations of hormones, leukocyte counts, microbial communities of the gut, immune responses, among many others, can be used to learn the structure of BNs. Additionally, poultry species are being studied under a variety of environmental conditions, such as heat/cold stress, different housing systems, or dietary regimens and/or supplements (Calefi et al., 2017; Lay et al., 2011; Luna et al., 2019; Pértille et al., 2020; Song et al., 2013; Sosnówka-Czajka et al., 2010), defining phenotypes associated with each particular condition (e.g., non-stressed vs stressed, non-supplemented vs supplemented). The possibilities are limitless and worth exploring.

8. Conclusions.

1. The genetics of stress of chickens represents a complex and challenging biological system, considering that experimental designs evaluated the combination of multiple factors, such as stress, domestication, and immune challenges.

2. Bioinformatic tools identified highly significant statistical genetic/epigenetic features, however, the lists of these genetic/epigenetic features are extensive, calling for additional tools to understand more in-depth the findings.

3. Bayesian networks represent a powerful mathematical tool that allowed the possibility to further explore the genetics/epigenetics of stress in chickens by learning functional and informative relationships between a given set of genes or methylated regions.

4. Dealing with a complex biological system required the exploration of the search space defined by the given set of variables to learn the structure of networks, the visualization of variability across searches, and the delineation of strategies and approaches to find the network that best fitted the data.

5. One gene, CARD19, and two differentially methylated regions, OCLN and ENS51236 displayed a direct link with the stress condition. These genetic/epigenetic features could be further explored as biomarkers of stress in chickens. Interestingly, ENS51236 is a DNA region whose function is still yet unknown, setting the stones for future studies.

6. The biology behind CARD19 and OCLN pointed towards the regulation of apoptotic process and the activation of the NF- κβ signaling pathways, and intracellular tight junctions and adhesion, respectively.

7. As stress signals, four Heat Shock Proteins were found to interact amongst themselves and with other four genes. This finding is relevant as the stress signals were initially identified in a small set of chickens and then searched for in other datasets, even though not all the datasets evaluated the effects of stress on the gene expression.

8. Our study highlights the power of Bayesian networks in knowledge discovery, not only in terms of the informative and functional interactions that were learnt, but also in the possibility to include genetic/epigenetic features whose functions are still yet unknown.

9. Interdisciplinary studies are needed to get further insights into complex biological systems. In this study, genetics, bioinformatics, statistics, and Bayesian networks were used in a series of steps to narrow the number of genetic/epigenetic features down, from hundreds of thousands to a small set. This way, it was possible to unravel key biological mechanisms and pathways as potential targets of future studies.

10. Considering that applying Bayesian network approaches involves a series of steps and making decisions, road maps were displayed so that other researchers can apply Bayesian network strategies on their own datasets. Consequently, they would be able to collect and combine different sources of data, properly discretize their data, explore the search space defined by the given set of variables, visualize the variability across searches, and finally, and build the network that best fits their data.

9. References.

Adelman, J. S., & Martin, L. B. (2009). Vertebrate sickness behaviors: Adaptive and integrated neuroendocrine immune responses. *Integrative and Comparative Biology, 49,* 202–214. https://doi.org/10.1093/icb/icp028

Affara, M., Sanders, D., Araki, H., Tamada, Y., Dunmore, B. J., Humphreys, S., … Charnock-Jones, D. S. (2013). Vasohibin-1 is identified as a master-regulator of endothelial cell apoptosis using gene network analysis. *BMC Genomics, 14,* 23. https://doi.org/10.1186/1471-2164-14-23

Aglipay, J. A., Lee, S. W., Okada, S., Fujiuchi, N., Ohtsuka, T., Kwak, J. C., … Ouchi, T. (2003). A member of the Pyrin family, IFI16, is a novel BRCA1-associated protein involved in the p53-mediated apoptosis pathway. *Oncogene, 22,* 8931-8938. https://doi.org/10.1038/sj.onc.1207057

Aglipay, J. A., Martin, S. A., Tawara, H., Lee, S. W., & Ouchi, T. (2006). ATM activation by ionizing radiation requires BRCA1-associated BAAT1. *Journal of Biological Chemistry, 281,* 9710-9718. https://doi.org/10.1074/jbc.M510332200

Agnvall, B., Jöngren, M., Strandberg, E., & Jensen, P. (2012). Heritability and genetic correlations of fear-related behaviour in red junglefowl-possible implications for early domestication. *PLoS ONE, 7,* e35162. https://doi.org/10.1371/journal.pone.0035162

Agrahari, R., Foroushani, A., Docking, T. R., Chang, L., Duns, G., Hudoba, M., … Zare, H. (2018). Applications of Bayesian network models in predicting types of hematological malignancies. *Scientific Reports, 8,* 6951. https://doi.org/10.1038/s41598-018-24758-5

Akbarian, A., Michiels, J., Degroote, J., Majdeddin, M., Golian, A., & De Smet, S. (2016). Association between heat stress and oxidative stress in poultry; mitochondrial dysfunction and dietary interventions with phytochemicals. *Journal of Animal Science and Biotechnology, 7,* 37. https://doi.org/10.1186/s40104-016-0097-5

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. D. (2010). Local

causal and markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research, 11,* 171-234.

Altan, O., Pabuçcuoğlu, A., Altan, A., Konyalioğlu, S., & Bayraktar, H. (2003). Effect of heat stress on oxidative stress, lipid peroxidation and some stress parameters in broilers. *British Poultry Science, 44,* 545–550. https://doi.org/10.1080/00071660310001618334

Alvarenga, R. R., Rodrigues, P. B., Zangeronimo, M. G., Freitas, R. T. F., Lima, R. R., Bertechini, A. G., & Fassani, E. J. (2011). Energetic values of feedstuffs for broilers determined with in vivo assays and prediction equations. *Animal Feed Science and Technology, 168,* 257-266. https://doi.org/10.1016/j.anifeedsci.2011.04.092

Alvarenga, T. C., Lima, R. R., Bueno Filho, J. S. S., Simão, S. D., Mariano, F. C. Q., Alvarenga, R. R., & Rodrigues, P. B. (2021). Application of Bayesian networks to the prediction of the AMEn: A new methodology in broiler nutrition. *Translational Animal Science, 5,* txaa215. https://doi.org/10.1093/tas/txaa215

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics, 25,* 25-29. https://doi.org/10.1038/75556

Ashley, N. T., & Demas, G. E. (2017). Neuroendocrine-immune circuits, phenotypes, and interactions. *Hormones and Behavior, 87,* 25–34. https://doi.org/10.1016/j.yhbeh.2016.10.004

Aslam, M. L., Bastiaansen, J. W. M., Elferink, M. G., Megens, H. J., Crooijmans, R. P. M. A., Blomberg, L. A., … Long, J. A. (2012). Whole genome SNP discovery and analysis of genetic diversity in Turkey (Meleagris gallopavo). *BMC Genomics, 13,* 391. https://doi.org/10.1186/1471-2164-13-391

Balov, N. (2013). A categorical network approach for discovering differentially expressed

regulations in cancer. *BMC Medical Genomics*, *6* (suppl. 3). https://doi.org/10.1186/1755-8794-6-S3-S1

Bélteky, J., Agnvall, B., Bektic, L., Höglund, A., Jensen, P., & Guerrero-Bosagna, C. (2018). Epigenetics and early domestication: Differences in hypothalamic DNA methylation between red junglefowl divergently selected for high or low fear of humans. *Genetics Selection Evolution, 50,* 13. https://doi.org/10.1186/s12711-018-0384-z

Bélteky, J., Agnvall, B., & Jensen, P. (2017). Gene expression of behaviorally relevant genes in the cerebral hemisphere changes after selection for tameness in Red Junglefowl. *PLoS ONE, 12,* e0177004. https://doi.org/10.1371/journal.pone.0177004

Bélteky, J., Agnvall, B., Johnsson, M., Wright, D., & Jensen, P. (2016). Domestication and tameness: Brain gene expression in red junglefowl selected for less fear of humans suggests effects on reproduction and immunology. *Royal Society Open Science, 3,* 160033. https://doi.org/10.1098/rsos.160033

Berkan Sesen, M., Nicholson, A. E., Banares-Alcantara, R., Kadir, T., & Brady, M. (2013). Bayesian networks for clinical decision support in lung cancer care. *PLoS ONE*, *8,* e82349. https://doi.org/10.1371/journal.pone.0082349

Bertin, J., Guo, Y., Wang, L., Srinivasula, S. M., Jacobson, M. D., Poyet, J. L., … Alnemri, E. S. (2000). CARD9 is a novel caspase recruitment domain-containing protein that interacts with BCL10/CLAP and activates NF-κB. *Journal of Biological Chemistry*, *275,* 41082-41086. https://doi.org/10.1074/jbc.C000726200

Bertin, J., Wang, L., Guo, Y., Jacobson, M. D., Poyet, J. L., Srinivasula, S. M., … Alnemri, E. S. (2001). CARD11 and CARD14 Are Novel Caspase Recruitment Domain (CARD)/Membrane-associated Guanylate Kinase (MAGUK) Family Members that Interact with BCL10 and Activate NF-κB. *Journal of Biological Chemistry*, *276,* 11877-11882. https://doi.org/10.1074/jbc.M010512200

Blanchard, M. A., Roskam, I., Mikolajczak, M., & Heeren, A. (2021). A network approach to parental burnout. *Child Abuse and Neglect, 111*, 104826. https://doi.org/10.1016/j.chiabu.2020.104826

Bowling, M., Forder, R., Hughes, R. J., Weaver, S., & Hynd, P. I. (2018). Effect of restricted feed intake in broiler breeder hens on their stress levels and the growth and immunology of their offspring. *Translational Animal Science, 2*, 263–271. https://doi.org/10.1093/tas/txy064

Burkholder, K. M., Thompson, K. L., Einstein, M. E., Applegate, T. J., & Patterson, J. A. (2008). Influence of stressors on normal intestinal microbiota, intestinal morphology, and susceptibility to Salmonella Enteritidis colonization in broilers. *Poultry Science, 87*, 1734-1741. https://doi.org/10.3382/ps.2008-00107

Burmester, T., Ebner, B., Weich, B., & Hankeln, T. (2002). Cytoglobin: A novel globin type ubiquitously expressed in vertebrate tissues. *Molecular Biology and Evolution, 19*, 416-421. https://doi.org/10.1093/oxfordjournals.molbev.a004096

Calefi, A. S., Da Silva Fonseca, J. G., Cohn, D. W. H., Honda, B. T. B., Costola-De-Souza, C., Tsugiyama, L. E., … Palermo-Neto, J. (2016). The gut-brain axis interactions during heat stress and avian necrotic enteritis. *Poultry Science, 95*, 1005-1014. https://doi.org/10.3382/ps/pew021

Calefi, A. S., Quinteiro-Filho, W. M., Ferreira, A. J. P., & Palermo-Neto, J. (2017). Neuroimmunomodulation and heat stress in poultry. *World's Poultry Science Journal, 73*, 493–504. https://doi.org/10.1017/S0043933917000472

Campbell, D. L. M., Hinch, G. N. Downing, J. A., & Lee, C (2016). Fear and coping styles of outdorr-preferring, moderate-outdoor and indoor-preferring free-range laying hens. *Applied Animal Behaviour Science, 185*, 73-77. https://doi.org/10.1016/j.applanim.2016.09.004

Cantet, J. M., Yu, Z., & Ríus, A. G. (2021). Heat stress-mediated activation of immune–inflammatory pathways. *Antibiotics*, *10*, 1285. https://doi.org/10.3390/antibiotics10111285

Carser, J. E., Quinn, J. E., Michie, C. O., O'Brien, E. J., McCluggage, W. G., Maxwell, P., … Harkin, D. P. (2011). BRCA1 is both a prognostic and predictive biomarker of response to chemotherapy in sporadic epithelial ovarian cancer. *Gynecologic Oncology*, *123*, 492-498. https://doi.org/10.1016/j.ygyno.2011.08.017

Cavalcante, R. G., Qin, T., & Sartor, M. A. (2018). Novel bioinformatics methods for toxicoepigenetics. In *Toxicoepigenetics: Core Principles and Applications*, 265-288. https://doi.org/10.1016/B978-0-12-812433-8.00012-5

Chen, D., Ning, F., Zhang, J., Tang, Y., & Teng, X. (2020). NF-κB pathway took part in the development of apoptosis mediated by miR-15a and oxidative stress via mitochondrial pathway in ammonia-treated chicken splenic lymphocytes. *Science of the Total Environment*, *729*, 139017. https://doi.org/10.1016/j.scitotenv.2020.139017

Chen, Y. C., Wheeler, T. A., & Kochenderfer, M. J. (2017). Learning discrete Bayesian networks from continuous data. *Journal of Artificial Intelligence Research*, *59*, 103-132. https://doi.org/10.1613/jair.5371

Cheng, H. W. (2010). Breeding of tomorrow's chickens to improve well-being. *Poultry Science*, *89*, 805-813. https://doi.org/10.3382/ps.2009-00361

Chickering, D. M. (1996). Learning Bayesian Networks is NP-Complete. In: Fisher D., Lenz HJ. (eds) Learning from Data. Lecture Notes in Statistics, 112. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-2404-4_12.

Ciaccio, M. F., Wagner, J. P., Chuu, C. P., Lauffenburger, D. A., & Jones, R. B. (2010). Systems analysis of EGF receptor signaling dynamics with microwestern arrays. *Nature Methods*, *7*, 148-155. https://doi.org/10.1038/nmeth.1418

Cockrem, J. F. (2007). Stress, corticosteron responses and avian personalities. *Journal of Ornithology, 148,* 169-178. https://doi.org/10.1007/s10336-007-0175-8

Comin, A., Jeremiasson, A., Kratzer, G., & Keeling, L. (2019). Revealing the structure of the associations between housing system, facilities, management and welfare of commercial laying hens using Additive Bayesian Networks. *Preventive Veterinary Medicine, 164,* 23-32. https://doi.org/10.1016/j.prevetmed.2019.01.004

Cooper, G. F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence, 42,* 393-405. https://doi.org/10.1016/0004-3702(90)90060-D

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics, 11,* 2463-2468. https://doi.org/10.1093/hmg/11.20.2463

Cruz-Ramírez, N., Acosta-Mesa, H. G., Carrillo-Calvet, H., Alonso Nava-Fernández, L., & Barrientos-Martínez, R. E. (2007). Diagnosis of breast cancer using Bayesian networks: A case study. *Computers in Biology and Medicine, 37,* 1553-1564. https://doi.org/10.1016/j.compbiomed.2007.02.003

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems, Complex Sy, 1695,* 1-9.

Davison, F., Kaspers, B., Schat, K. A., & Kaiser, P. (2014). *Avian immunology*. Academic Press.

de Campos, L. M., & Huete, J. F. (2000). Approximating Causal Orderings for Bayesian Networks Using Genetic Algorithms and Simulated Annealing. *Proceedings of the Eighth Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, 2000,* 211-219.

De Kloet, E. R. (2003). Hormones, brain and stress. *Endocrine Regulations, 37,* 51–68.

de Haas, E. N., Kops, M. S., Bolhuis, J. E., Groothuis, T. G. G., Ellen, E. D., & Rodenburg, T. B. (2012). The relation between fearfulness in young and stress-response in adult laying hens: on individual and group level. *Physiology and Behaviour, 107*, 433-439. https://doi.org/10.1016/j.physbeh.2012.08.002

Deist, M. S., Gallardo, R. A., Bunn, D. A., Dekkers, J. C. M., Zhou, H., & Lamont, S. J. (2017). Resistant and susceptible chicken lines show distinctive responses to Newcastle disease virus infection in the lung transcriptome. *BMC Genomics*, *18*, 989. https://doi.org/10.1186/s12864-017-4380-4

Dhabhar, F S, & McEwen, B. S. (1997). Acute stress enhances while chronic stress suppresses cell-mediated immunity in vivo: a potential role for leukocyte trafficking. *Brain, Behavior, and Immunity*, *11*, 286–306. https://doi.org/10.1006/brbi.1997.0508

Dhabhar, Firdaus S. (2009). Enhancing versus suppressive effects of stress on immune function: implications for immunoprotection and immunopathology. *Neuroimmunomodulation*, *16*, 300–317. https://doi.org/10.1159/000216188

Dickens, M. J., & Romero, L. M. (2013). A consensus endocrine profile for chronically stressed wild animals does not exist. *General and Comparative Endocrinology*, *15*, 177-189. https://doi.org/10.1016/j.ygcen.2013.06.014

Dixon, L. M. (2020). Slow and steady wins the race: The behaviour and welfare of commercial faster growing broiler breeds compared to a commercial slower growing breed. *PLoS ONE, 15*, e0231006. https://doi.org/10.1371/journal.pone.0231006

Djebbari, A., & Quackenbush, J. (2008). Seeded Bayesian Networks: Constructing genetic networks from microarray data. *BMC Systems Biology*, *2*, 57. https://doi.org/10.1186/1752-0509-2-57

Dohms, J. E., & Metz, A. (1991). Stress--mechanisms of immunosuppression. *Veterinary Immunology and Immunopathology*, *30*, 89–109. https://doi.org/10.1016/0165-

2427(91)90011-Z

Domingos, P., & Pazzani, M. (1996). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *Machine Learning*.

Dong, H., Lin, H., Jiao, H. C., Song, Z. G., Zhao, J. P., & Jiang, K. J. (2007). Altered development and protein metabolism in skeletal muscles of broiler chickens (Gallus gallus domesticus) by corticosterone. *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology, 147*, 189-195. https://doi.org/10.1016/j.cbpa.2006.12.034

Du, Y., Xu, J. T., Jin, H. N., Zhao, R., Zhao, D., Du, S. H., … Wang, Q. (2017). Increased cerebral expressions of MMPs, CLDN5, OCLN, ZO1 and AQPs are associated with brain edema following fatal heat stroke. *Scientific Reports, 7*, 1691. https://doi.org/10.1038/s41598-017-01923-w

Ducatelle, R., Goossens, E., De Meyer, F., Eeckhaut, V., Antonissen, G., Haesebrouck, F., & Van Immerseel, F. (2018). Biomarkers for monitoring intestinal health in poultry: Present status and future perspectives. *Veterinary Research, 49*, 43. https://doi.org/10.1186/s13567-018-0538-6

Elfwing, M., Nätt, D., Goerlich-Jansson, V. C., Persson, M., Hjelm, J., & Jensen, P. (2015). Early stress causes sex-specific, life-long changes in behaviour, levels of gonadal hormones, and gene expression in chickens. *PLoS ONE, 10*, e0125808. https://doi.org/10.1371/journal.pone.0125808

Ericsson, M., Henriksen, R., Bélteky, J., Sundman, A. S., Shionoya, K., & Jensen, P. (2016). Long-term and transgenerational effects of stress experienced during different life phases in chickens (Gallus gallus). *PLoS ONE, 11*, e0153879. https://doi.org/10.1371/journal.pone.0153879

Ericsson, M., & Jensen, P. (2016). Domestication and ontogeny effects on the stress response

in young chickens (Gallus gallus). *Scientific Reports, 6*, 35818. https://doi.org/10.1038/srep35818

Fallahshahroudi, A., Løtvedt, P., Bélteky, J., Altimiras, J., & Jensen, P. (2019). Changes in pituitary gene expression may underlie multiple domesticated traits in chickens. *Heredity, 122*, 195-204. https://doi.org/10.1038/s41437-018-0092-z

Fallahsharoudi, A., de Kock, N., Johnsson, M., Bektic, L., Ubhayasekera, S. J. K. A., Bergquist, J., … Jensen, P. (2017). QTL mapping of stress related gene expression in a cross between domesticated chickens and ancestral red junglefowl. *Molecular and Cellular Endocrinology, 5*, 52-58. https://doi.org/10.1016/j.mce.2017.02.010

Fallahsharoudi, A., De Kock, N., Johnsson, M., Ubhayasekera, S. J. K. A., Bergquist, J., Wright, D., & Jensen, P. (2015). Domestication effects on stress induced steroid secretion and adrenal gene expression in chickens. *Scientific Reports, 5*, 15345. https://doi.org/10.1038/srep15345

Fallahsharoudi, A., Kock, N. de, Johnsson, M., Bektic, L., Ubhayasekera, S. J. K. A., Bergquist, J., … Jensen, P. (2017). Genetic and targeted eQTL mapping reveals strong candidate genes modulating the stress response during chicken domestication. *G3: Genes, Genomes, Genetics*, *7*, 497-504. https://doi.org/10.1534/g3.116.037721

Faverial, J., Cornet, D., Paul, J., & Sierra, J. (2016). Multivariate analysis of the determinants of the end-product quality of manure-based composts and vermicomposts using Bayesian network modelling. *PLoS ONE*, *11*, e0157884. https://doi.org/10.1371/journal.pone.0157884

Felipe, V. P. S., Silva, M. A., Valente, B. D., & Rosa, G. J. M. (2014). Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. *Poultry Science, 94*, 772-780. https://doi.org/10.3382/ps/pev031

Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, *7*, 1728-1740. https://doi.org/10.1038/nprot.2012.101

Fernandez, M. E., Labaque, M. C., Orso, G., Marin, R. H., & Kembro, J. M. (2021). Short- and long-term dynamics of the physiological and behavioral response to heat stress and thymol supplementation in Japanese quail. *Journal of Thermal Biology*, *97*, 102876. https://doi.org/10.1016/j.jtherbio.2021.102876

Furuse, M., & Tsukita, S. (2006). Claudins in occluding junctions of humans and flies. *Trends in Cell Biology*, *16*, 181-188. https://doi.org/10.1016/j.tcb.2006.02.006

Gallup, G. G. (1977). Tonic Immobility: the Role of Fear and Predation. *The Psychological Record*, *27*, 41-61. https://doi.org/10.1007/bf03394432

Gasparino, E., Voltolini, D. M., Del Vesco, A. P., Guimarães, S. E. F., do Nascimento, C. S., & de Oliveira Neto, A. R. (2013). IGF-I, GHR and UCP mRNA expression in the liver and muscle of high- and low-feed-efficiency laying Japanese quail at different environmental temperatures. *Livestock Science*, *157*, 339–344. https://doi.org/10.1016/j.livsci.2013.06.028

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*, 307-315. https://doi.org/10.1093/bioinformatics/btg405

Giayetto, O., Videla, E. A., Chacana, P., Jaime, C. E., Marin, R. H., & Nazar, F. N. (2020). Modulating offspring responses: concerted effects of stress and immunogenic challenge in the parental generation. *Journal of Experimental Biology*, *223*, jeb219386. https://doi.org/10.1242/jeb.219386

Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J., & Hayes, B. J. (2016). Genetics of complex traits: Prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proceedings of the Royal Society B: Biological*

Sciences, 283, 20160569. https://doi.org/10.1098/rspb.2016.0569

Goel, A., Ncho, C. M., & Choi, Y. H. (2021). Regulation of gene expression in chickens by heat stress. *Journal of Animal Science and Biotechnology*, *12*, 11. https://doi.org/10.1186/s40104-020-00523-5

Goerlich, V. C., Nätt, D., Elfwing, M., Macdonald, B., & Jensen, P. (2012). Transgenerational effects of early experience on behavioral, hormonal and gene expression responses to acute stress in the precocial chicken. *Hormones and Behavior, 61*, 711-718. https://doi.org/10.1016/j.yhbeh.2012.03.006

Greene, C. S., Tan, J., Ung, M., Moore, J. H., & Cheng, C. (2014). Big data bioinformatics. *Journal of Cellular Physiology, 229*, 1896-1900. https://doi.org/10.1002/jcp.24662

Guo, L., Zhao, G., Xu, J. R., Kistler, H. C., Gao, L., & Ma, L. J. (2016). Compartmentalized gene regulatory network of the pathogenic fungus Fusarium graminearum. *The New Phytologist*, *211*, 527-541. https://doi.org/10.1111/nph.13912

Guo, Y., Jiang, R., Su, A., Tian, H., Zhang, Y., Li, W., … Kang, X. (2020). Identification of genes related to effects of stress on immune function in the spleen in a chicken stress model using transcriptome analysis. *Molecular Immunology*, *124*, 180-189. https://doi.org/10.1016/j.molimm.2020.06.004

Guzmán, D. A., Flesia, A. G., Aon, M. A., Pellegrini, S., Marin, R. H., & Kembro, J. M. (2017). The fractal organization of ultradian rhythms in avian behavior. *Scientific Reports, 7*, 684. https://doi.org/10.1038/s41598-017-00743-2

Hahsler, M., Grün, B., & Hornik, K. (2005). arules - A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, *14*, 1-25. https://doi.org/10.18637/jss.v014.i15

Hartnack, S., Odoch, T., Kratzer, G., Furrer, R., Wasteson, Y., L'Abée-Lund, T. M., & Skjerve, E. (2019). Additive Bayesian networks for antimicrobial resistance and

potential risk factors in non-typhoidal Salmonella isolates from layer hens in Uganda. *BMC Veterinary Research*, *15*, 212. https://doi.org/10.1186/s12917-019-1965-y

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, *20*, 197-243. https://doi.org/10.1023/A:1022623210503

Hedlund, L., Palazon, T., & Jensen, P. (2021). Stress during commercial hatchery processing induces long-time negative cognitive judgement bias in chickens. *Animals*, *11*, 1083. https://doi.org/10.3390/ani11041083

Henriksen, R., Rettenbacher, S., & Groothuis, T. G. G. (2011). Prenatal stress in birds: Pathways, effects, function and perspectives. *Neuroscience and Biobehavioral Reviews*, *35*, 1484-1501. https://doi.org/10.1016/j.neubiorev.2011.04.010

Hernández, J., & Thompson, I. M. (2004). Prostate-specific antigen: A review of the validation of the most commonly used cancer biomarker. *Cancer*, *101*, 894-904. https://doi.org/10.1002/cncr.20480

Hidano, A., Yamamoto, T., Hayama, Y., Muroga, N., Kobayashi, S., Nishida, T., & Tsutsui, T. (2015). Unraveling antimicrobial resistance genes and phenotype patterns among Enterococcus faecalis isolated from retail chicken products in Japan. *PLoS ONE*, *10*, e0121189. https://doi.org/10.1371/journal.pone.0121189

Hirakawa, R., Nurjanah, S., Furukawa, K., Murai, A., Kikusato, M., Nochi, T., & Toyomizu, M. (2020). Heat Stress Causes Immune Abnormalities via Massive Damage to Effect Proliferation and Differentiation of Lymphocytes in Broiler Chickens. *Frontiers in Veterinary Science*, *7*, 46. https://doi.org/10.3389/fvets.2020.00046

Hofmann, K. (1999). The modular nature of apoptotic signaling proteins. *Cellular and Molecular Life Sciences*, *55*, 1113-1128. https://doi.org/10.1007/s000180050361

Honda, B. T. B., Calefi, A. S., Costola-De-Souza, C., Quinteiro-Filho, W. M., Da Silva

Fonseca, J. G., De Paula, V. F., & Palermo-Neto, J. (2015). Effects of heat stress on peripheral T and B lymphocyte profiles and IgG and IgM serum levels in broiler chickens vaccinated for Newcastle disease virus. *Poultry Science, 94*, 2375-2381. https://doi.org/10.3382/ps/pev192

Howard, D. M., Hall, L. S., Hafferty, J. D., Zeng, Y., Adams, M. J., Clarke, T. K., … McIntosh, A. M. (2017). Genome-wide haplotype-based association analysis of major depressive disorder in Generation Scotland and UK Biobank. *Translational Psychiatry, 7*, 1263. https://doi.org/10.1038/s41398-017-0010-9

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*, 44-57. https://doi.org/10.1038/nprot.2008.211

Huff, G. R., Huff, W. E., & Balog, J. M. (2005). Stress response differences and disease susceptibility reflected by heterophil to lymphocyte ratio in turkeys selected for increased body weight. *Poultry Science, 84*, 709–717. doi: 10.1093/ps/84.5.709.

Jackson, S., & Diamond, J. (1996). Metabolic and digestive responses to artificial selection in chickens. *Evolution, 50*, 1638-1650. https://doi.org/10.1111/j.1558-5646.1996.tb03936.x

Jones, R. B., Satterlee, D. G., & Ryder, F. H. (1992). Fear and distress in Japanese quail chicks of two lines genetically selected for low or high adrenocortical response to immobilization stress. *Hormones and Behavior, 26*, 385–393.

Jung, J., Eggleton, P., Robinson, A., Wang, J., Gutowski, N., Holley, J., … Michalak, M. (2018). Calnexin is necessary for T cell transmigration into the central nervous system. *JCI Insight, 3*, e98410. https://doi.org/10.1172/jci.insight.98410

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research, 40*(D1), D109-D114. https://doi.org/10.1093/nar/gkr988

Katajamaa, R., & Jensen, P. (2020). Selection for reduced fear in red junglefowl changes brain composition and affects fear memory. *Royal Society Open Science, 7*, 200628. https://doi.org/10.1098/rsos.200628

Kelemen, A., Zhou, H., Lawhead, P., & Liang, Y. (2003). Naive Bayesian Classifier for Microarray Data. *Proceedings of the International Joint Conference on Neural Networks, 3,* 1769-1773. https://doi.org/10.1109/ijcnn.2003.1223675

Kembro, J. M., Satterlee, D. G., Schmidt, J. B., Perillo, M. A., & Marin, R. H. (2008). Open-field temporal pattern of ambulation in Japanese quail genetically selected for contrasting adrenocortical responsiveness to brief manual restraint. *Poultry Science, 87*, 2186–2195. https://doi.org/10.3382/ps.2008-00108

Koussounadis, A., Langdon, S. P., Harrison, D. J., & Smith, V. A. (2014). Chemotherapy-induced dynamic gene expression changes in vivo are prognostic in ovarian cancer. *British Journal of Cancer*, *110,* 2975-2984. https://doi.org/10.1038/bjc.2014.258

Kuchipudi, S. V., Tellabati, M., Sebastian, S., Londt, B. Z., Jansen, C., Vervelde, L., … Chang, K. C. (2014). Highly pathogenic avian influenza virus infection in chickens but not ducks is associated with elevated host immune and pro-inflammatory responses. *Veterinary Research, 45*, 118. https://doi.org/10.1186/s13567-014-0118-3

Kuenzel, W. J., & Jurkevich, A. (2010). Molecular neuroendocrine events during stress in poultry. *Poultry Science*, *89*(4), 832–840. https://doi.org/10.3382/ps.2009-00376

Kugelstadt, D., Haberkamp, M., Hankeln, T., & Burmester, T. (2004). Neuroglobin, cytoglobin, and a novel, eye-specific globin from chicken. *Biochemical and Biophysical Research Communications, 325,* 719-725. https://doi.org/10.1016/j.bbrc.2004.10.080

Kumuthini, J., Bessant, C., Wilson, P., & Crowther, D. (2007). Gene regulatory network of human adipocyte differentiation. *BMC Systems Biology, 1*, P57. https://doi.org/10.1186/1752-0509-1-s1-p57

Kyrimi, E., McLachlan, S., Dube, K., Neves, M. R., Fahmi, A., & Fenton, N. (2021). A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future. *Artificial Intelligence in Medicine*, *117*, 102108 https://doi.org/10.1016/j.artmed.2021.102108

Lábaque, M. C., Kembro, J. M., Luna, A., & Marin, R. H. (2013). Effects of thymol feed supplementation on female Japanese quail (Coturnix coturnix) behavioral fear response. *Animal Feed Science and Technology*, *183*, 67–72. https://doi.org/10.1016/j.anifeedsci.2013.04.018

Lai, E., Teodoro, T., & Volchuk, A. (2007). Endoplasmic reticulum stress: Signaling the unfolded protein response. *Physiology*, *22*, 193-201. https://doi.org/10.1152/physiol.00050.2006

Lam, W., & Bacchus, F. (1994). Learning bayesian belief networks: an approach based on the mdl principle. *Computational Intelligence*, *10*, 269-293. https://doi.org/10.1111/j.1467-8640.1994.tb00166.x

Lamon, S. J., Cobl, D. J., Bjorkquis, A., Rothschil, M. F., Persi, M., Ashwel, C., & Schmid, C. (2014). Genomics of heat stress in chickens. *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production*, *12*, 11. https://doi.org/10.1186/s40104-020-00523-5

Lara, L., & Rostagno, M. (2013). Impact of Heat Stress on Poultry Production. *Animals*, *3*, 356–369. https://doi.org/10.3390/ani3020356

Lay, D. C., Fulton, R. M., Hester, P. Y., Karcher, D. M., Kjaer, J. B., Mench, J. A., … Porter, R. E. (2011). Hen welfare in different housing systems. *Poultry Science*, *90*, 278–294. https://doi.org/10.3382/ps.2010-00962

Li, F., Hu, G., Zhang, H., Wang, S., Wang, Z., & Li, H. (2013). Epistatic effects on abdominal fat content in chickens: Results from a genome-wide SNP-SNP interaction

analysis. *PLoS ONE*, *8*, e81520. https://doi.org/10.1371/journal.pone.0081520

Li, H., Wu, G., Zhang, J., & Yang, N. (2010). Identification of the heart-type fatty acid-binding protein as a major gene for chicken fatty acid metabolism by bayesian network analysis. *Poultry Science*, *89*, 1825-1833. https://doi.org/10.3382/ps.2010-00699

Li, W., & Gao, F. B. (2003). Actin filament-stabilizing protein tropomyosin regulates the size of dendritic fields. *Journal of Neuroscience*, *23*, 6171-6175. https://doi.org/10.1523/jneurosci.23-15-06171.2003

Lin, H., Sui, S. J., Jiao, H. C., Buyse, J., & Decuypere, E. (2006). Impaired development of broiler chickens by stress mimicked by corticosterone exposure. *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology*, *143*, 400–405. https://doi.org/10.1016/j.cbpa.2005.12.030

Lin, H., Decuypere, E., & Buyse, J. (2006). Acute heat stress induces oxidative stress in broiler chickens. *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology*, *144*, 11-17. https://doi.org/10.1016/j.cbpa.2006.01.032

Liu, J., & Bo, S. (2011). Naive Bayesian classifier based on genetic simulated annealing algorithm. *Procedia Engineering*, *23*, 504-509. https://doi.org/10.1016/j.proeng.2011.11.2538

Long, N., Gianola, D., Rosa, G. J., Weigel, K. A., & Avendaño, S. (2009). Comparison of classification methods for detecting associations between SNPs and chick mortality. *Genetics Selection Evolution*, *41*, 18. https://doi.org/10.1186/1297-9686-41-18

Løtvedt, P., Fallahshahroudi, A., Bektic, L., Altimiras, J., & Jensen, P. (2017). Chicken domestication changes expression of stress-related genes in brain, pituitary and adrenals. *Neurobiology of Stress*, *7*, 113-121. https://doi.org/10.1016/j.ynstr.2017.08.002

Luna, A., Tarifa, M. F., Fernandez, M. E., Caliva, J. M., Pellegrini, S., Zygadlo, J. A., & Marin, R. H. (2019). Thymol, alpha tocopherol, and ascorbyl palmitate supplementation

as growth enhancers for broiler chickens. *Poultry Science, 98,* 1012-1016. https://doi.org/10.3382/ps/pey362

Luo, Q. B., Song, X. Y., Ji, C. L., Zhang, X. Q., & Zhang, D. X. (2014). Exploring the molecular mechanism of acute heat stress exposure in broiler chickens using gene expression profiling. *Gene, 546,* 200-205. https://doi.org/10.1016/j.gene.2014.06.017

Lupo, C., Le Bouquin, S., Balaine, L., Michel, V., Péraste, J., Petetin, I., … Chauvin, C. (2013). Bayesian network as an aid for Food Chain Information use for meat inspection. *Preventive Veterinary Medicine, 109,* 25-36. https://doi.org/10.1016/j.prevetmed.2012.09.004

Ma, Y., Gu, L., Yang, L., Sun, C., Xie, S., Fang, C., … Li, S. (2018). Identifying artificial selection signals in the chicken genome. *PLoS ONE, 13,* e0196215. https://doi.org/10.1371/journal.pone.0196215

Malheiros, R. D., Moraes, V. M. B., Collin, A., Decuypere, E., & Buyse, J. (2003). Free diet selection by broilers as influenced by dietary macronutrient ratio and corticosterone supplementation. 1. Diet selection, organ weights, and plasma metabolites. *Poultry Science, 83,* 123-131. https://doi.org/10.1093/ps/82.1.123

Mantione, K. J., Kream, R. M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J. M., & Stefano, G. B. (2014). Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Medical Science Monitor Basic Research, 20,* 138–141. https://doi.org/10.12659/MSMBR.892101

Mashaly, M. M., Hendricks, G. L., Kalama, M. A., Gehad, A. E., Abbas, A. O., & Patterson, P. H. (2004). Effect of heat stress on production parameters and immune responses of commercial laying hens. *Poultry Science, 83,* 889–894. https://doi.org/10.1093/ps/83.6.889

Matthäus, F., Smith, V. A., & Gebicke-Haerter, P. J. (2010). Some Useful Mathematical

Tools to Transform Microarray Data into Interactive Molecular Networks. In *Systems Biology in Psychiatric Research: From High-Throughput Data to Mathematical Modeling*. Wiley-Blackwell. https://doi.org/10.1002/9783527630271.ch13

McEver, R. P., & Cummings, R. D. (1997). Role of PSGL-1 binding to selectins in leukocyte recruitment. *Journal of Clinical Investigation*, *100*, S97-103. https://doi.org/10.1172/jci119556

McEwen, B. S., Biron, C. A., Brunson, K. W., Bulloch, K., Chambers, W. H., Dhabhar, F. S., … Weiss, J. M. (1997). The role of adrenocorticoids as modulators of immune function in health and disease: Neural, endocrine and immune interactions. *Brain Research Reviews*, *23*, 79-133. https://doi.org/10.1016/S0165-0173(96)00012-4

McNally, R. J., Heeren, A., & Robinaugh, D. J. (2017). A Bayesian network analysis of posttraumatic stress disorder symptoms in adults reporting childhood sexual abuse. *European Journal of Psychotraumatology*, *8*(sup3), 1341276. https://doi.org/10.1080/20008198.2017.1341276

Meier, M., Kossakowski, J. J., Jones, P. J., Kay, B., Riemann, B. C., & McNally, R. J. (2020). Obsessive–compulsive symptoms in eating disorders: A network investigation. *International Journal of Eating Disorders*, *53*, 362-371. https://doi.org/10.1002/eat.23196

Mellor, D. J., & Beausoleil, N. J. (2015). Extending the "Five Domains" model for animal welfare assessment to incorporate positive welfare states. *Animal Welfare, 24*, 241-253. https://doi.org/10.7120/09627286.24.3.241

Mellor, D. J. (2016). Updating animalwelfare thinking: Moving beyond the "five freedoms" towards "A lifeworth living." *Animals, 6*. https://doi.org/10.3390/ani6030021

Mignon-Grasteau, S., Boissy, A., Bouix, J., Faure, J. M., Fisher, A. D., Hinch, G. N., … Beaumont, C. (2005). Genetics of adaptation and domestication in livestock. *Livestock*

Production Science, 93, 3-14. https://doi.org/10.1016/j.livprodsci.2004.11.001

Miller, K. A., & Mench, J. A. (2006). Differential effects of 4 types of environmental enrichment on aggressive pecking, feather pecking, feather loss, food wastage and productivity in Japanese quail. *British Poultry Science*, *47*, 646–658. https://doi.org/10.1080/00071660601084333

Milns, I., Beale, C. M., & Anne Smith, V. (2010). Revealing ecological networks using Bayesian network inference algorithms. *Ecology, 91*, 1892-1899. https://doi.org/10.1890/09-0731.1

Mitchell, E. G., Wallace, M. I., Smith, V. A., Wiesenthal, A. A., & Brierley, A. S. (2021). Bayesian Network Analysis reveals resilience of the jellyfish Aurelia aurita to an Irish Sea regime shift. *Scientific Reports*, *11*, 3707. https://doi.org/10.1038/s41598-021-82825-w

Moekti, G. R. (2020). Industrial livestock production: A review on advantages and disadvantages. *IOP Conference Series: Earth and Environmental Science, 492*, 012094. https://doi.org/10.1088/1755-1315/492/1/012094

Mumma, J. O., Thaxton, J. P., Vizzier-Thaxton, Y., & Dodson, W. L. (2006). Physiological stress in laying hens. *Poultry Science, 85*, 761–769.

Nagarajan, R., Scutari, M., & Lèbre, S. (2013a). Bayesian Networks in R: with Applications in Systems Biology. In *Bayesian Networks in R: with Applications in Systems Biology*. https://doi.org/10.1007/978-1-4614-6446-4

Nagarajan, R., Scutari, M., & Lèbre, S. (2013b). Bayesian Networks in R. In *Bayesian Networks in R*. https://doi.org/10.1007/978-1-4614-6446-4

Nazar, F. N., Videla, E. A., Fernandez, M. E., Labaque, M. C., & Marin, R. H. (2018). Insights into thermal stress in Japanese quail (Coturnix coturnix): dynamics of immunoendocrine and biochemical responses during and after chronic exposure. *Stress,*

21, 257-266. https://doi.org/10.1080/10253890.2018.1442430

Nazar, F.N., Videla, E. A., & Marin, R. H. (2018). Thymol supplementation effects on adrenocortical, immune and biochemical variables recovery in Japanese quail after exposure to chronic heat stress. *Animal, 13*, 318-325. https://doi.org/10.1017/S175173111800157X

Nazar, F N, Magnoli, A. P., Dalcero, A. M., & Marin, R. H. (2012). Effect of feed contamination with aflatoxin B1 and administration of exogenous corticosterone on Japanese quail biochemical and immunological parameters. *Poultry Science*, *91*, 47–54. https://doi.org/10.3382/ps.2011-01658

Nazar, F.N., Barrios, B. E., Kaiser, P., Marin, R. H., & Correa, S. G. (2015). Immune neuroendocrine phenotypes in Coturnix coturnix: Do avian species show LEWIS/FISCHER-like profiles? *PLoS ONE*, *10*, e0120712. https://doi.org/10.1371/journal.pone.0120712

Nazar, F.N., Estevez, I., Correa, S. G., & Marin, R. H. (2017). Stress Induced Polarization of Immune-Neuroendocrine Phenotypes in Gallus gallus. *Scientific Reports*, *7*, 8102. https://doi.org/10.1038/s41598-017-08733-0

Needham, C. J., Bradford, J. R., Bulpitt, A. J., & Westhead, D. R. (2007). A primer on learning in Bayesian networks for computational biology. *PLoS Computational Biology, 3*, e129. https://doi.org/10.1371/journal.pcbi.0030129

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 69*, 026113. https://doi.org/10.1103/PhysRevE.69.026113

Niazi, U., Geyer, K. K., Vickers, M. J., Hoffmann, K. F., & Swain, M. T. (2016). DISMISS: Detection of stranded methylation in MeDIP-Seq data. *BMC Bioinformatics*, *17*, 295. https://doi.org/10.1186/s12859-016-1158-7

Noguera, J. C., Aira, M., Pérez-Losada, M., Domínguez, J., & Velando, A. (2018). Glucocorticoids modulate gastrointestinal microbiome in a wild bird. *Royal Society Open Science, 5*, 171743. https://doi.org/10.1098/rsos.171743

Nojavan A., F., Qian, S. S., & Stow, C. A. (2017). Comparative analysis of discretization methods in Bayesian networks. *Environmental Modelling and Software, 87*, 64-71. https://doi.org/10.1016/j.envsoft.2016.10.007

Okada, S., & Ouchi, T. (2003). Cell cycle differences in DNA damage-induced BRCA1 phosphorylation affect its subcellular localization. *Journal of Biological Chemistry, 278*, 2015-2020. https://doi.org/10.1074/jbc.M208685200

Ouyang, Y. Bin, & Moore, K. L. (1998). Molecular cloning and expression of human and mouse tyrosylprotein sulfotransferase-2 and a tyrosylprotein sulfotransferase homologue in Caenorhabditis elegans. *Journal of Biological Chemistry, 273*, 24770-24774. https://doi.org/10.1074/jbc.273.38.24770

Park, W., Srikanth, K., Lim, D., Park, M., Hur, T., Kemp, S., … Park, J. E. (2019). Comparative transcriptome analysis of Ethiopian indigenous chickens from low and high altitudes under heat stress condition reveals differential immune response. *Animal Genetics, 50*, 42-53. https://doi.org/10.1111/age.12740

Parsons, D. J., Orton, T. G., D'Souza, J., Moore, A., Jones, R., & Dodd, C. E. R. (2005). A comparison of three modelling approaches for quantitative risk assessment using the case study of Salmonella spp. in poultry meat. *International Journal of Food Microbiology, 98*, 35-51. https://doi.org/10.1016/j.ijfoodmicro.2004.05.005

Pearl, J. (1988). BOOK - Pearl - Probabilistic reasoning in intelligent systems: : networks of plausible inference (Morgan kaufmann series in representation and reasoning). *Morgan Kaufmann Publishers, San Mateo, Calif.*

Perini, F., Cendron, F., Rovelli, G., Castellini, C., Cassandro, M., & Lasagna, E. (2021).

Emerging genetic tools to investigate molecular pathways related to heat stress in chickens: A review. *Animals, 11*, 46. https://doi.org/10.3390/ani11010046

Pértille, F., Brantsæter, M., Nordgreen, J., Coutinho, L. L., Janczak, A. M., Jensen, P., & Guerrero-Bosagna, C. (2017). DNA methylation profiles in red blood cells of adult hens correlate with their rearing conditions. *Journal of Experimental Biology, 220*, 3579-3587. https://doi.org/10.1242/jeb.157891

Pértille, F., Ibelli, A. M. G., Sharif, M. El, Poleti, M. D., Fröhlich, A. S., Rezaei, S., … Coutinho, L. L. (2020). Putative Epigenetic Biomarkers of Stress in Red Blood Cells of Chickens Reared Across Different Biomes. *Frontiers in Genetics*, *11*, 508809. https://doi.org/10.3389/fgene.2020.508809

Peters, J., Bellet, B. W., Jones, P. J., Wu, G. W. Y., Wang, L., & McNally, R. J. (2021). Posttraumatic stress or posttraumatic growth? Using network analysis to explore the relationships between coping styles and trauma outcomes. *Journal of Anxiety Disorders, 78*, 102359. https://doi.org/10.1016/j.janxdis.2021.102359

Pfau, M. L., & Russo, S. J. (2015). Peripheral and central mechanisms of stress resilience. *Neurobiology of Stress, 1*, 66-79. https://doi.org/10.1016/j.ynstr.2014.09.004

Philippe, F. X., Mahmoudi, Y., Cinq-Mars, D., Lefrançois, M., Moula, N., Palacios, J., … Godbout, S. (2020). Comparison of egg production, quality and composition in three production systems for laying hens. *Livestock Science, 232*, 103917. https://doi.org/10.1016/j.livsci.2020.103917

Price, E. O. (1999). Behavioral development in animals undergoing domestication. *Applied Animal Behaviour Science, 65*, 245-271. https://doi.org/10.1016/S0168-1591(99)00087-8

Pusch, E. A., Bentz, A. B., Becker, D. J., & Navara, K. J. (2018) Behavioral phenotype predicts physiological responses to chronic stress in proactive and reactive birds.

General and Comparative Endrocrinology, 255, 71-77. https://doi.org/10.1016/j.ygcen.2017.10.008

Qiu, L., Chang, G., Li, Z., Bi, Y., Liu, X., & Chen, G. (2018). Comprehensive transcriptome analysis reveals competing endogenous RNA networks during avian leukosis virus, subgroup J-induced tumorigenesis in chickens. *Frontiers in Physiology, 9*, 996. https://doi.org/10.3389/fphys.2018.00996

Quinteiro-Filho, W. M., Gomes, A. V. S., Pinheiro, M. L., Ribeiro, A., Ferraz-de-Paula, V., Astolfi-Ferreira, C. S., … Palermo-Neto, J. (2012). Heat stress impairs performance and induces intestinal inflammation in broiler chickens infected with Salmonella Enteritidis. *Avian Pathology, 41*, 421–427. https://doi.org/10.1080/03079457.2012.709315

Quinteiro-Filho, W. M., Ribeiro, A., Ferraz-De-Paula, V., Pinheiro, M. L., Sakai, M., Sá, L. R. M., … Palermo-Neto, J. (2010). Heat stress impairs performance parameters, induces intestinal injury, and decreases macrophage activity in broiler chickens. *Poultry Science, 89*, 1905–1914. https://doi.org/10.3382/ps.2010-00812

R Core Team. (2021). R core team (2021). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL Http://Www. R-Project. Org*.

Rao, M. S., Van Vleet, T. R., Ciurlionis, R., Buck, W. R., Mittelstadt, S. W., Blomme, E. A. G., & Liguori, M. J. (2019). Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. *Frontiers in Genetics, 9*, 636. https://doi.org/10.3389/fgene.2018.00636

Redmond, S. B., Chuammitri, P., Andreasen, C. B., Palić, D., & Lamont, S. J. (2009). Chicken heterophils from commercially selected and non-selected genetic lines express cytokines differently after in vitro exposure to Salmonella enteritidis. *Veterinary Immunology and Immunopathology, 132*, 129–134.

https://doi.org/10.1016/j.vetimm.2009.05.010

Ren, Z., Wang, Y., Deng, H., Deng, Y., Deng, J., Zuo, Z., … Shen, L. (2015). Deoxynivalenol induces apoptosis in chicken splenic lymphocytes via the reactive oxygen species-mediated mitochondrial pathway. *Environmental Toxicology and Pharmacology, 39,* 339-346. https://doi.org/10.1016/j.etap.2014.11.028

Renaudeau, D., Collin, A., Yahav, S., de Basilio, V., Gourdine, J. L., & Collier, R. J. (2012). Adaptation to hot climate and strategies to alleviate heat stress in livestock production. *Animal, 6,* 707–728. https://doi.org/10.1017/s1751731111002448.

Rezaei S., Uffenorde J., Gimm O., Hosseinpour Feizi M.A., Miemczyk S., Lehmann Coutinho L., ... Pértille, F. (2022). GBS-MeDIP: A protocol for parallel identification of genetic and epigenetic variation in the same reduced fraction of genomes across individuals. *STAR Protocols, 3,* 101202.

Ricard, N., Scott, R. P., Booth, C. J., Velazquez, H., Cilfone, N. A., Baylon, J. L., … Simons, M. (2019). Endothelial ERK1/2 signaling maintains integrity of the quiescent endothelium. *Journal of Experimental Medicine, 216,* 1874-1890. https://doi.org/10.1084/jem.20182151

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43,* e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26,* 139-140. https://doi.org/10.1093/bioinformatics/btp616

Rodgers, R. F., DuBois, R., Thiebaut, S., Jaussent, I., Maimoun, L., Seneque, M., … Guillaume, S. (2019). Structural Differences in Eating Disorder Psychopathology After History of Childhood Abuse: Insights From a Bayesian Network Analysis. *Journal of*

Abnormal Psychology, 128, 795-805. https://doi.org/10.1037/abn0000470

Romero, L. M., Dickens, M. J., & Cyr, N. E. (2009). The reactive scope model - A new model integrating homeostasis, allostasis, and stress. *Hormones and Behavior, 55,* 375–389. https://doi.org/10.1016/j.yhbeh.2008.12.009

Rosales, A. G. (1994). Managing Stress in Broiler Breeders: A Review. *The Journal of Applied Poultry Research, 3,* 199-207. https://doi.org/10.1093/japr/3.2.199

Rosen, E. M., Fan, S., Pestell, R. G., & Goldberg, I. D. (2003). BRCA1 gene in breast cancer. *Journal of Cellular Physiology, 196,* 19-41. https://doi.org/10.1002/jcp.10257

Ross, M., Rausch, Q., Vandenberg, B., & Mason, G. (2020). Hens with benefits: Can environmental enrichment make chickens more resilient to stress? *Physiology and Behavior, 226,* 113077. https://doi.org/10.1016/j.physbeh.2020.113077

Saelao, P., Wang, Y., Gallardo, R. A., Lamont, S. J., Dekkers, J. M., Kelly, T., & Zhou, H. (2018). Novel insights into the host immune response of chicken Harderian gland tissue during Newcastle disease virus infection and heat treatment. *BMC Veterinary Research, 14,* 280. https://doi.org/10.1186/s12917-018-1583-0

Sapolsky, R., Romero, L., & Munck, A. (2000). How Do Glucocorticoids Influence Stress Responses ? Integrating Permissive, Suppressive, Stimulatory, and Preparative Actions. *Endocrine Reviews, 21,* 55–89. https://doi.org/10.1210/er.21.1.55

Scanes, C. G. (2016). Biology of stress in poultry with emphasis on glucocorticoids and the heterophil to lymphocyte ratio. *Poultry Science, 95,* 2208–2215. https://doi.org/10.3382/ps/pew137

Schmidt, M., Gerlach, F., Avivi, A., Laufs, T., Wystub, S., Simpson, J. C., … Burmester, T. (2004). Cytoglobin Is a Respiratory Protein in Connective Tissue and Neurons, Which Is Up-regulated by Hypoxia. *Journal of Biological Chemistry, 279,* 8063-8069. https://doi.org/10.1074/jbc.M310540200

Schuster, M. K., Schmierer, B., Shkumatava, A., & Kuchler, K. (2004). Activin A and Follicle-Stimulating Hormone Control Tight Junctions in Avian Granulosa Cells by Regulating Occludin Expression. *Biology of Reproduction, 70*, 1493-1499. https://doi.org/10.1095/biolreprod.103.024331

Scully, R., Chen, J., Ochs, R. L., Keegan, K., Hoekstra, M., Feunteun, J., & Livingston, D. M. (1997). Dynamic changes of BRCA1 subnuclear location and phosphorylation state are initiated by DNA damage. *Cell, 90*, 425-435. https://doi.org/10.1016/S0092-8674(00)80503-6

Scutari M (2010). Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, *35*, 1–22. https://doi.org/10.18637/jss.v035.i03

Sekelova, Z., Stepanova, H., Polansky, O., Varmuzova, K., Faldynova, M., Fedr, R., … Vlasatikova, L. (2017). Differential protein expression in chicken macrophages and heterophils in vivo following infection with Salmonella Enteritidis. *Veterinary Research*, *48*, 35. https://doi.org/10.1186/s13567-017-0439-0

Selye, H. (1950). Stress and the general adaptation syndrome. *British Medical Journal*, *1*, 1383–1392. https://doi.org/10.1136/bmj.2.4670.104-a

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*, 1135–1145.

Shi, D., Bai, L., Qu, Q., Zhou, S., Yang, M., Guo, S., … Liu, C. (2019). Impact of gut microbiota structure in heat-stressed broilers. *Poultry Science, 98*, 2405-2413. https://doi.org/10.3382/ps/pez026

Shinde, S., Zerbs, S., Collart, F. R., Cumming, J. R., Noirot, P., & Larsen, P. E. (2019). Pseudomonas fluorescens increases mycorrhization and modulates expression of antifungal defense response genes in roots of aspen seedlings. *BMC Plant Biology, 19*, 4. https://doi.org/10.1186/s12870-018-1610-0

Shini, S., Kaiser, P., Shini, A., & Bryden, W. L. (2008). Biological response of chickens (Gallus gallus domesticus) induced by corticosterone and a bacterial endotoxin. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology, 149*, 324-333. https://doi.org/10.1016/j.cbpb.2007.10.003

Siegel, H. S. (1971). Adrenals, Stress and the Environment. *World's Poultry Science Journal, 27*, 327-349. https://doi.org/10.1079/wps19710038

Siegel, H. S. (1980). Physiological Stress in Birds. *BioScience, 30*, 529–534. https://doi.org/10.2307/1307973

Siegel, H. S. (1995). Gordon Memorial Lecture. Stress, strains and resistance. *British Poultry Science*, *36*, 3–22. https://doi.org/10.1080/00071669508417748

Skinner, M. K., Manikkam, M., & Guerrero-Bosagna, C. (2010). Epigenetic transgenerational actions of environmental factors in disease etiology. *Trends in Endocrinology and Metabolism, 21*, 214-222. https://doi.org/10.1016/j.tem.2009.12.007

Smith, V. A. (2010). Revealing structure of complex biological systems using bayesian networks. In *Network Science: Complexity in Nature and Technology*. https://doi.org/10.1007/978-1-84996-396-1_9

Smith V.A., Yu J, Smulders T.V., Hartemink A.J., & Jarvis E.D (2006). Computational inference of neural information flow networks. *PLoS Computational Biology, 2*, e161. https://doi.org/10.1371/journal.pcbi.0020161

Soleimani, A. F., Zulkifli, I., Omar, A. R., & Raha, A. R. (2011). Physiological responses of 3 chicken breeds to acute heat stress. *Poultry Science, 90,* 1435–1440. https://doi.org/10.3382/ps.2011-01381

Soltani, M., & Omid, M. (2015). Detection of poultry egg freshness by dielectric spectroscopy and machine learning techniques. *LWT - Food Science and Technology, 62*, 1034-1042. https://doi.org/10.1016/j.lwt.2015.02.019

Song, J., Jiao, L. F. F., Xiao, K., Luan, Z. S. S., Hu, C. H. H., Shi, B., & Zhan, X. A. A. (2013). Cello-oligosaccharide ameliorates heat stress-induced impairment of intestinal microflora, morphology and barrier integrity in broilers. *Animal Feed Science and Technology*, *185*, 175–181. https://doi.org/10.1016/j.anifeedsci.2013.08.001

Sosnówka-Czajka, E., Herbut, E., & Skomorucha, I. (2010). Effect of different housing systems on productivity and welfare of laying hens. *Annals of Animal Science, 1-*, 349-360.

Sperandio, M., Gleissner, C. A., & Ley, K. (2009). Glycosylation in immune cell trafficking. *Immunological Reviews*, *230*, 97-113. https://doi.org/10.1111/j.1600-065X.2009.00795.x

Spradling, K. D., McDaniel, A. E., Lohi, J., & Pilcher, B. K. (2001). Epsin 3 Is a Novel Extracellular Matrix-induced Transcript Specific to Wounded Epithelia. *Journal of Biological Chemistry*, *276*, 29257-29267. https://doi.org/10.1074/jbc.M101663200

Stephens, C. S., & Johnson, P. A. (2017). Occludin expression and regulation in small follicles of the layer and broiler breeder hen. *General and Comparative Endocrinology*, *248*, 106-113. https://doi.org/10.1016/j.ygcen.2017.02.010

Tarca, A. L., Carey, V. J., Chen, X. wen, Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology*, *3*, e116. https://doi.org/10.1371/journal.pcbi.0030116

Tixier-Boichard, M., Bed'Hom, B., & Rognon, X. (2011). Chicken domestication: From archeology to genomics. *Comptes Rendus - Biologies*, *334*, 197-204. https://doi.org/10.1016/j.crvi.2010.12.012

Tkac, J., Gajdosova, V., Hroncekova, S., Bertok, T., Hires, M., Jane, E., … Kasak, P. (2019). Prostate-specific antigen glycoprofiling as diagnostic and prognostic biomarker of prostate cancer. *Interface Focus*, *9*, 20180077. https://doi.org/10.1098/rsfs.2018.0077

Van Goor, A., Ashwell, C. M., Persia, M. E., Rothschild, M. F., Schmidt, C. J., & Lamont, S. J. (2017). Unique genetic responses revealed in RNAseq of the spleen of chickens stimulated with lipopolysaccharide and short-term heat. *PLoS ONE*, *12*, e0171414. https://doi.org/10.1371/journal.pone.0171414

Vandana, G. D., Sejian, V., Lees, A. M., Pragna, P., Silpa, M. V., & Maloney, S. K. (2021). Heat stress and poultry production: impact and amelioration. *International Journal of Biometeorology*, *65*, 163-179. https://doi.org/10.1007/s00484-020-02023-7

Vasilescu, C., Dragomir, M., Tanase, M., Giza, D., Purnichescu-Purtan, R., Chen, M., … Calin, G. A. (2017). Circulating miRNAs in sepsis - A network under attack: An in-silico prediction of the potential existence of miRNA sponges in sepsis. *PLoS ONE*, *12*, e0183334. https://doi.org/10.1371/journal.pone.0183334

Vogogias, A., Kennedy, J., Archambault, D., Bach, B., Anne Smith, V., & Currant, H. (2018). Bayespiles: Visualisation support for Bayesian network structure learning. *ACM Transactions on Intelligent Systems and Technology*, *10*, 5-23. https://doi.org/10.1145/3230623

Wang, L., Guo, Y., Huang, W. J., Ke, X., Poyet, J. L., Manji, G. A., … Bertin, J. (2001). CARD10 Is a Novel Caspase Recruitment Domain/Membrane-associated Guanylate Kinase Family Member That Interacts with BCL10 and Activates NF-κB. *Journal of Biological Chemistry*, *276*, 21405-21409. https://doi.org/10.1074/jbc.M102488200

Wang, Y., Lupiani, B., Reddy, S. M., Lamont, S. J., & Zhou, H. (2014). RNA-seq analysis revealed novel genes and signaling pathway associated with disease resistance to avian influenza virus infection in chickens. *Poultry Science*, *93*, 485-493. https://doi.org/10.3382/ps.2013-03557

Wang, Y., Saelao, P., Chanthavixay, K., Gallardo, R., Bunn, D., Lamont, S. J., … Zhou, H. (2018). Physiological responses to heat stress in two genetically distinct chicken inbred

lines. *Poultry Science, 97*, 770-780. https://doi.org/10.3382/ps/pex363

Webster, A. J. (2001). Farm animal welfare: the five freedoms and the free market. *Veterinary Journal (London, England : 1997)*, *161*, 229–237. https://doi.org/10.1053/tvjl.2000.0563

Webster, J. (2016). Animal welfare: Freedoms, dominions and "A life worth living." *Animals, 6*, 35. https://doi.org/10.3390/ani6060035

Webster Marketon, J. I., & Glaser, R. (2008). Stress hormones and immune function. *Cellular Immunology*, *252*, 16–26. https://doi.org/10.1016/j.cellimm.2007.09.006

Wickramasuriya, S. S., Park, I., Lee, K., Lee, Y., Kim, W. H., Nam, H., & Lillehoj, H. S. (2022). Role of Physiology, Immunity, Microbiota, and Infectious Diseases in the Gut Health of Poultry. *Vaccines*, *10*, 172. https://doi.org/10.3390/vaccines10020172

Windhorst, H. W. (2006). Changes in poultry production and trade worldwide. *World's Poultry Science Journal, 62*, 585-602. https://doi.org/10.1079/WPS2006114

Xia, W. X., Yu, Q., Li, G. H., Liu, Y. W., Xiao, F. H., Yang, L. Q., … Kong, Q. P. (2019). Identification of four hub genes associated with adrenocortical carcinoma progression by WGCNA. *PeerJ*, *14*, e6555. https://doi.org/10.7717/peerj.6555

Xie, J., Tang, L., Lu, L., Zhang, L., Xi, L., Liu, H. C., … Luo, X. (2014). Differential expression of heat shock transcription factors and heat shock proteins after acute and chronic heat stress in laying chickens (Gallus gallus). *PLoS ONE*, *9*, 3102204. https://doi.org/10.1371/journal.pone.0102204

Xie, X., Cho, B., & Fischer, J. A. (2012). Drosophila Epsin's role in Notch ligand cells requires three Epsin protein functions: The lipid binding function of the ENTH domain, a single Ubiquitin interaction motif, and a subset of the C-terminal protein binding modules. *Developmental Biology*, *363*, 399-412. https://doi.org/10.1016/j.ydbio.2012.01.004

Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., & Jarvis, E. D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, *20*, 3594-3603. https://doi.org/10.1093/bioinformatics/bth448

Yu, Y., Hoffhines, A. J., Moore, K. L., & Leary, J. A. (2007). Determination of the sites of tyrosine O-sulfation in peptides and proteins. *Nature Methods*, *4*, 583-588. https://doi.org/10.1038/nmeth1056

Zahoor, I., De Koning, D. J., & Hocking, P. M. (2017). Transcriptional profile of breast muscle in heat stressed layers is similar to that of broiler chickens at control temperature. *Genetics Selection Evolution, 49,* 69. https://doi.org/10.1186/s12711-017-0346-x

Zhang, B., Shao, Y., Liu, D., Yin, P., Guo, Y., & Yuan, J. (2012). Zinc prevents Salmonella enterica serovar Typhimurium-induced loss of intestinal mucosal barrier function in broiler chickens. *Avian Pathology, 41*, 361-367. https://doi.org/10.1080/03079457.2012.692155

Zhang, H., Shen, L. Y., Xu, Z. C., Kramer, L. M., Yu, J. Q., Zhang, X. Y., … Li, H. (2020). Haplotype-based genome-wide association studies for carcass and growth traits in chicken. *Poultry Science*, *99,* 2349-2361. https://doi.org/10.1016/j.psj.2020.01.009

Zhang, H., Yu, J. Q., Yang, L. L., Kramer, L. M., Zhang, X. Y., Na, W., … Li, H. (2017). Identification of genome-wide SNP-SNP interactions associated with important traits in chicken. *BMC Genomics, 18,* 892. https://doi.org/10.1186/s12864-017-4252-y

Zidar, J., Sorato, E., Malmqvist, A. M., Jansson, E., Rosher, C., Jensen, P., … Løvlie, H. (2017). Early experience affects adult personality in the red junglefowl: A role for cognitive stimulation? *Behavioural Processes*, *134*, 78–86. https://doi.org/10.1016/j.beproc.2016.06.003

Zilliox, M. J., & Irizarry, R. A. (2007). A gene expression bar code for microarray data. *Nature Methods*, *4*, 911-913. https://doi.org/10.1038/nmeth1102

Zulkifli, I., Al-Aqil, A., Omar, A. R., Sazili, A. Q., & Rajion, M. A. (2009). Crating and heat stress influence blood parameters and heat shock protein 70 expression in broiler chickens showing short or long tonic immobility reactions. *Poultry Science*, *88*, 471–476. https://doi.org/10.3382/ps.2008-00287