Genetic variation in the chicken genome: insights in selection

Martin Elferink

Thesis committee

Thesis supervisor

Prof. dr. M.A.M. Groenen
Personal chair at Animal Breeding and Genomics Centre
Wageningen University

Thesis co-supervisor

Dr. R.P.M.A. Crooijmans
Assistant Professor at Animal Breeding and Genomics Centre
Wageningen University

Other Members

Prof. dr. ir. F.P.M. Govers
Wageningen University

Prof. dr. J.H.S.G.M. de Jong
Wageningen University

Prof. dr. M. Pérez-Enciso
Universitat Autònoma de Barcelona, Bellaterra, Spain

Dr. M. Tixier-Boichard
Ministère de l'Enseignement supérieur et de la Recherche, Paris, France

This research was conducted under the auspices of the Wageningen Institute of Animal Sciences (WIAS) graduate school

Genetic variation in the chicken genome: insights in selection

Martin Elferink

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 20 June 2011
at 4 p.m. in the Aula.

Martin Gerhard Elferink
Genetic variation in the chicken genome: insights in selection,
164 pages

Thesis, Wageningen University, Wageningen, NL (2011)
With references, with summaries in Dutch and English

ISBN 978-90-8585-920-8

Contents

1

General introduction

1.1 Introduction

Taxonomy of *Gallus*

Birds and mammals evolved separately from their common ancestor approximately 310 million years ago [1]. Four species of the genus *Gallus* are known to modern ornithology and currently exist as wild populations [2] (Figure 1.1). The *Gallus lafayetii* (Sri Lanka Junglefowl) lives in Sri Lanka, the *Gallus sonneratii* (Grey Junglefowl) in western and southern India, the *Gallus varius* (Green Junglefowl) in Indonesia (Java and neighboring islands), and the *Gallus gallus* (Red Junglefowl) in a large part of Asia, including Northeast India, Southern China, and Southeast Asia. Based on observations described by Darwin, crosses between these species result in infertile offspring [3]. However, additional hybridization experiments performed in the mid-19th century between *Gallus gallus* and *Gallus sonneratii* [4] and *Gallus gallus* and *Gallus varius* [5] resulted in fertile offspring.

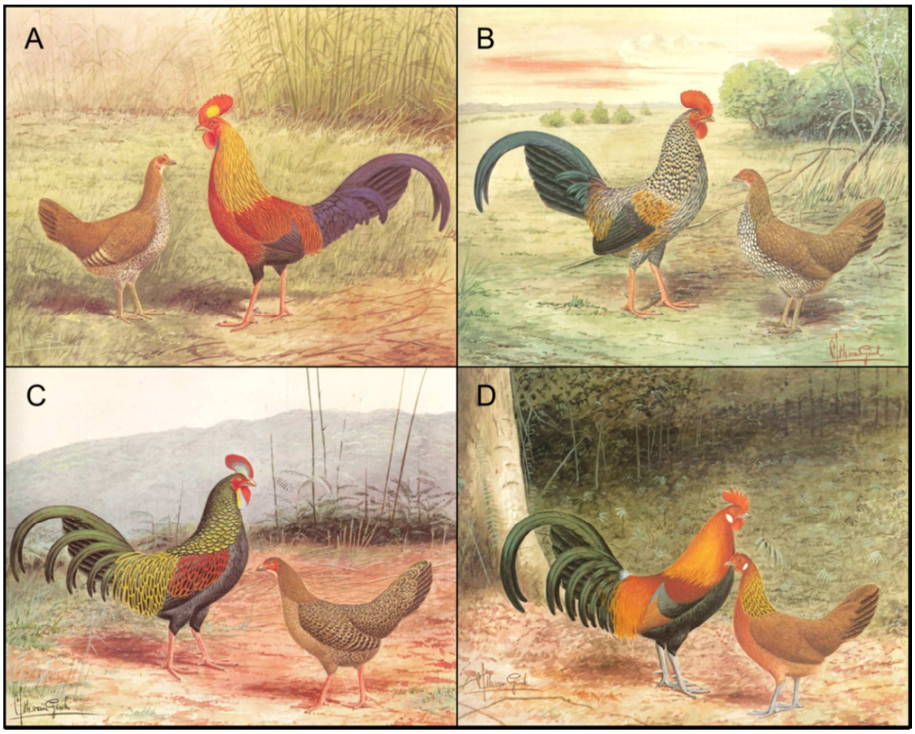


Figure 1.1 The four *Gallus* species. A) *Gallus lafayetii*. B) *Gallus sonneratii*. C) *Gallus varius*. D) *Gallus gallus*. Source: the pictures are provided by Stichting Fonds voor Pluimveebelangen in the Netherlands.

The widely spread *Gallus gallus* (G.g.) consists of five subspecies; the *G.g. gallus* in Indochina, the *G.g. bankiva* in Java (Indonesia), the *G.g. jabouillei* in Vietnam, the *G.g. murghi* in India, and the *G.g. spadiceus* in Burma. There are no subspecies known for the *G. lafayetii, G. sonneratii, and G. varius*.

The domesticated chicken

Charles Darwin proposed that the domesticated chicken originated exclusively from *Gallus gallus* [3]. The domestic chicken is, therefore, classified as *G.g. domesticus*. Although the single-origin was supported by many studies (e.g. [6,7,8,9,10], it was debated by others [11,12]. Molecular genetic evidence supports multiple instances and multiple regions of domestication of the chicken from Red Junglefowl. Moreover, recent evidence supports genetic contributions from other Junglefowl species to current domesticated chickens. For instance, the yellow skin locus present in several domestic chicken breeds most likely originated from the *Gallus sonneratii* [13]. Archeological findings, moreover, suggest that multiple domestication events were involved in the establishment of the domesticated chicken [14,15,16]. Archeological findings suggest that one of the domestication events in chickens occurred 8000 BP in Southeast Asia [17].

The chicken may initially not have been domesticated as a new food resource, but mainly for cultural reasons such as religion, decoration, and cock fighting [2]. The domesticated chicken gradually spread to other regions of the world. Each of these regions had their own culture and environment, thereby influencing the evolution of the domesticated chicken. In the 19th century the so called 'hen craze' in Europe and the Americas also had a large influence on the evolution of the chicken [18]. Chicken became very popular for hobby purposes to royalty and upper classes and most breeds currently in existence in Europe and the Americas were developed in that period [2].

The commercial chicken breeds

Although selective breeding of chickens as a food resource has been documented to occur by the time of the Roman Empire [2], the strongest artificial selection most likely took place in the 20th century by commercial breeding companies. Specialized lines, intensely selected on either growth traits (meat production) or reproductive traits (egg-laying) led to a massive increase in these production traits [19,20,21]. At the present time there are essentially four different commercial breeds; white egg-layers, brown egg-layers, broiler (meat-type chicken) sire lines and broiler dam lines (Figure 1.2). All white egg-layers are based on the white leghorn breed that originated from Livorno in Tuscany, Italy [2,22]. Brown egg-

layers are mainly based on Rhode Island Red and White Plymouth rock breeds [22]. Within the broiler lines, there are two distinct lines known as sire and dam lines. The sire lines are specifically selected on growth performance traits, while the dam lines are selected on growth and fertility traits. This distinct separation is necessary because of the negative pleiotropic effect between growth and fertility. Broiler sire lines are based on the Cornish breed, while broiler dam lines are based on several different breeds such as the White Plymouth rock, Barred Plymouth rock, and New Hampshire breeds [22]. All commercial lines are essentially closed and no gene-flow occurs between commercial and non-commercial breeds [22].

The spectacular progresses in both egg and meat production traits of commercial chickens are, however, also associated with an increased occurrence of undesirable traits such as reduced fertility [23], reduced resistance to infectious disease [24], skeletal deformities [25], congenital disorders [20], osteoporosis [26], and the pulmonary hypertension syndrome [21,27,28,29].

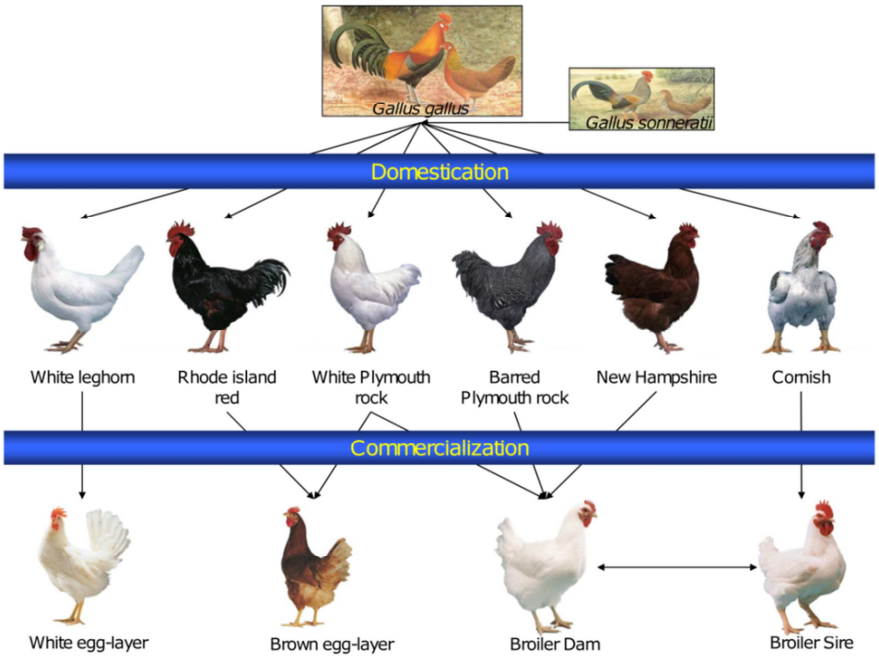


Figure 1.2 Origin of the four commercial breeds. Source: the pictures of the *G. gallus, G. sonneratii*, White leghorn, Rhode Island Red, White Plymouth rock, Barred Plymouth rock, New Hampshire, and Cornish are provided by Stichting Fonds voor Pluimveebelangen in the Netherlands. The White egg-layer and Brown egg-layer pictures are provided by ISA BV, Boxmeer, the Netherlands. The Broiler Dam and Sire pictures are provided by Cobb-Vantress Inc., Arkansas, USA.

The need for improvement of chicken breeds

In the past half-century the worldwide food production has increased tremendously [30] (Table 1.1). With a massive production increase in the last half-century (over 1000%) the chicken currently provides more than a quarter of the worldwide meat production. Furthermore, the worldwide egg production is still solely provided by the chicken (Table 1.1). The global human population will continue to grow and therefore an increased food production is needed in the near future [30]. Further improvements in meat and egg production traits in the chicken will contribute to this future need. However, to improve on these production traits animal breeders need to focus on the increased occurrence of undesired traits as these will increase production costs and reduce production progress [20]. Undesired traits should, moreover, be reduced to meet future demands for food safety (i.e. decreased use of chemicals and antibiotics to treat diseases) and to improve animal welfare.

Table 1.1 Production quantity of meat and eggs (tonnes) in 1961 and 2009.

Meat	1961	2009	Production increase (%)
Pig meat	24798970 (34.7)	106069157 (37.7)	428
Chicken meat	7555887 (10.6)	79595987 (28.3)	1053
Cattle meat	27684560 (38.8)	61837770 (22.0)	223
Sheep meat	4930305 (6.9)	8109219 (2.9)	164
Turkey meat	900630 (1.3)	5319748 (1.9)	591
Goat meat	1101886 (1.5)	4938655 (1.8)	448
Total meat	71410007 (100)	281559122 (100)	394
Eggs			
Hen eggs	14409313 (95.2)	62426378 (92.6)	433
Other birds	725645 (4.8)	4981371 (7.4)	686
Total eggs	15134959 (100)	67407749 (100)	445

For the meat production, only the major livestock species are included in this table. Meat total includes all species contribution to meat production. The percentage of total production is given between brackets. Information obtained from FAOSTAT webpage (http://faostat.fao.org/).

Although traditional phenotypic based breeding has proven to be successful, progress can be slow for traits that can only be measured later in life, in one sex, after slaughter, or if phenotypic measurements are expensive, for instance by requiring experimental facilities to challenge disease development [31]. The identification of genetic variation underlying the production and disease traits will replace the need for phenotypic measurements in each generation, thereby reducing costs, and enhancing genetical progress.

Once the genetic variation is identified, marker assisted selection (MAS) [32,33,34,35] or genomic selection (GS) [36] can aid in the improvement of production traits and disease resistance. The detection of causative variant underlying traits will, moreover, provide valuable insight in the biological mechanisms (for instance genes and biochemical pathways) of production and disease traits. Insights in the biological mechanisms could assist in further improvement of production due to more effective disease treatments and improved nutrients or housing conditions. The biological insights will, moreover, provide information for disease treatments in other livestock species or humans.

Genomic resources in the chicken

Besides an important livestock species, the chicken is also an important model species for biological research [37]. The chicken is a model species for classical genetics, developmental biology, immunology and evolutionary biology [38]. A large number of genomic resources have been developed for the chicken. These resources include linkage maps (e.g. [39,40,41,42,43,44,45]), radiation hybrid maps (e.g.[46,47,48]), EST and cDNA libraries [49,50], BAC libraries (e.g. [51,52]), clone based physical maps [53,54,55] and a large number of genetic markers such as SNPs and microsatellites (e.g. [56,57,58]). The chicken was the first livestock species to have its genome completely sequenced and annotated [59]. The total size of the chicken genome is approximately 1,1 giga base pairs, which is roughly one third of mammalian genomes [59].

The available linkage, radiation and physical maps assisted in the genome assembly, and the EST and cDNA libraries assisted in the gene annotation of the genome. In parallel with the sequencing project, more than 2.8 million SNPs were detected between the sequenced RJF and a single individual from Silkie, broiler and white egg-layer breeds [58]. In November 2010, the number of SNPs identified in chicken increased to more than 11 million, thereby being the third species after human and mouse with the highest number of SNPs in the dbSNP database (www.ncbi.nlm.nih.gov/projects/SNP/).

Despite the availability of a reference genome, further improvements are still needed for the chicken genome assembly. Within the first genome assembly, 11 out of the 40 chromosomes were underrepresented or completely missing. Moreover, it was estimated that 5-10% of all genes in the chicken genome are truncated or completely missing in the first assembly [59]. In May 2006 the second build (WASHUC2) of the chicken genome was released (www.ensembl.org). Although improvements were made in this assembly, the same 11 chromosomes remained underrepresented or were still completely missing. Moreover, numerous false segmental duplications in the chicken reference genome caused by mis-assembly were identified [60]. The third assembly of the chicken genome is expected to be released in 2011. In this third build the false segmental duplications have been corrected and, additional sequencing efforts by advanced sequencing technologies have resulted in closing sequence gaps. Both, the high resolution linkage map described in chapter 2 of this thesis [45], and the linkage map described by Groenen *et al.* [44] were used to improve this new assembly.
Besides assisting in the sequence assembly of genomes, linkage maps are important to study recombination rates and recombination hotspots within the genome. Accurate detection of recombination rates are, for instance, important for coalescence simulations used in hitch-hiking mapping [61]. In linkage maps the distance between markers is based on the recombination frequency between marker pairs rather than on the physical distance in base pairs. The distance in genetic maps is measured in centiMorgans (cM). The first linkage maps in the chicken were based on random amplified polymorphic DNA, chicken repeat element 1, restriction fragment length polymorphism, amplified fragment length polymorphism and microsatellite markers (e.g. [39,41,42,43]). The most recent linkage maps are mainly based on SNP markers [44,45].

Monogenic and polygenic traits

In genetics a distinction is made between traits that are either caused by a single gene variant (monogenic trait) or influenced by multiple genes and environmental factors (polygenic, quantitative or complex trait). A monogenic trait is caused by a single gene, can be either recessive or dominant, and follows patterns of Mendelian inheritance. The causative variant in monogenic diseases therefore explains 100% of the phenotypic variation of a trait.
There have been several successes in the detection of the causative variant in monogenic traits in livestock species. In chicken for instance, several variants were identified that are involved in plumage color [62,63,64] and fishy odor in eggs [65]. Moreover, several causative variants have been described in several other livestock

species, including various coat color traits in pig, cattle, dog and horse (e.g. [66,67,68,69,70], muscular hypertrophy in cattle [71] and sheep [72], malignant hyperthermia in pig [73], glycogen content in skeletal muscle in pig [74], and narcolepsy in dogs [75].

The causative variants underlying most of these monogenic traits were detected in a similar two tier approach. The first step includes linkage analysis (LA) to map the trait on a chromosome or linkage group. LA makes use of the fact that regions on the genome co-segregate with the trait phenotype in pedigrees over multiple generations or in independent families [76]. Association is found based on this co-segregation (linkage) between the phenotype and alleles of genetic markers such as RFLP, microsatellite, or SNPs. The second step generally includes fine-mapping of the associated region, comparative mapping to identify genes within the associated region, followed by the sequencing of (functional) candidate genes to identify causative variants.

Although several successes were obtained with LA, there are also severe limitations of this method. The confident intervals of the mapped QTLs are generally large and usually contain many candidate genes thereby hampering the identification of underlying causative variants. Moreover, the power to detect variants that have modest or low phenotypic effect is low unless an unrealistic large number of families is used for the analysis [77]. Most variants involved in polygenic traits have, however, low or modest phenotypic effect [78]. Therefore, LA is not particularly usefull to detect variants in polygenic traits [79].

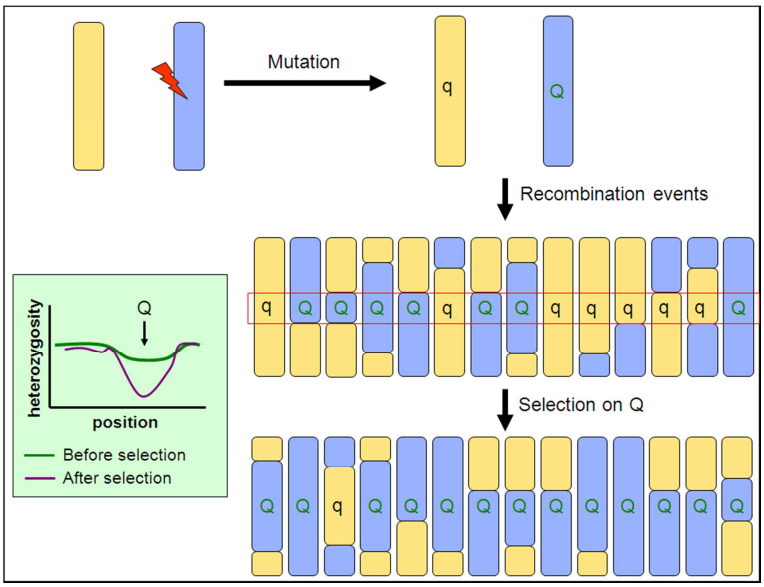
A polygenic (or quantitative) trait has a complex background and is influenced by environmental factors and multiple quantitative trait loci (QTL) each having a phenotypic effect on the trait phenotype [80]. Polygenic traits do not follow patterns of Mendelian inheritance. These complex traits include, for instance, growth, fertility, behavior, and diseases such as hypertension, schizophrenia and diabetes. Because most important production and disease traits in livestock species have a polygenic background [80], improved mapping methods are required to identify QTLs. With the increased genomic resources and rapid developments in high-throughput genotyping and sequencing techniques, new methods based on genome-wide marker assays such as linkage disequilibrium and hitch-hiking mapping have become available.

Linkage disequilibrium and genome-wide assays

With the increased genomic resources in chicken and the rapid developments in high-throughput genotyping techniques, genotyping assays including tens of thousands of SNPs became available for mapping studies [44,81]. Due to the existence of linkage disequilibrium (LD), only a limited number of genetic markers are needed to capture all genetic variation of the genome. LD refers to the non-random association of alleles at different loci [82] (Box 1). Markers are in LD when the combination of alleles occurs more frequently than would be expected based on their individual allele frequency. If two markers are in perfect LD, one marker can be used to capture the genetic information of the other. A group of two or more genetic markers in high LD is also known as a haplotype block, or haploblock. A haplotype is the combination of two or more alleles on the same chromosome that tend to be inherited together. Careful selection of markers within a haploblock – known as tag-SNPs – can provide information of all haplotypes within the block [83]. Thus, with a limited set of tag-SNPs it is possible to capture all genetic variation within the genome.

Recombination and mutation rates can vary throughout the genome, resulting in low LD at regions containing hotspots of recombination or regions prone to mutations (i.e. regions with many CpG dinucleotides) [76]. The variability in recombination is particularly strong in the chicken genome as it consists of macro- and microchromosomes that show distinct differences. Microchromosomes have a much higher recombination rate compared to macrochromosomes [44,84]. Megens *et al.* [85] detected reduced LD in microchromosomes and showed that this reduction was almost completely explained by differences in recombination rate. It was also shown that LD differed greatly between different commercial breeds in chicken [86]. In order to design marker assays that accurately cover the whole genome it is, therefore, essential to understand the LD structure and recombination patterns of the genome [85]. This understanding is needed to interpret results obtained from these genome-wide assays. However, due to limited knowledge on the LD structure of the genome the first genome-wide assays developed in livestock species were not based on tag-SNPs, but on common SNPs that are evenly spaced throughout the genome. The capacity of these first genome-wide assays, furthermore, was limited and did not allow for complete coverage of the genome.

Box 1. Linkage disequilibrium and genetic hitch-hiking.



Linkage disequilibrium

Consider a hypothetical population in which two chromosomes are segregating (yellow and blue) (top left). If a mutation occurs (Q) in one chromosome (blue), the wildtype allele (q) will remain in the other chromosome (yellow) (top right). This new mutation will create linkage disequilibrium because all alleles specific for the blue chromosome will be exclusively found in combination with Q and the allele specifically found in the yellow chromosome will be exclusively found with q. However, due to recombination events over multiple generations, recombinants between the blue and yellow chromosomes will occur (middle right) and LD will decrease. In the red boxed chromosomal segments, however, Q remains exclusively found in combination with the blue and q with the yellow chromosome segments. These blue and yellow segments in the red box are identical-by-descent. Besides recombination events, LD can also be decreased due to the occurrence of new mutations. LD can, however, also increase as a result of selection (either natural or artificial), genetic drift (i.e. loss of haplotypes), population growth, admixture of populations, population structure, and gene conversions [105].

Genetic hitch-hiking

Selection on Q will lead to an increased frequency of this allele in the population (before selection there are 7 Q and 7 q alleles, after selection 13 Q and only one q). When an allele increases in frequency, the heterozygosity at that particular locus will reduce. However, due to the existence of LD, the nucleotide diversity (allelic heterozygosity) will also be reduced at surrounding loci (green inset on the left). This loss of nucleotide diversity at and near the selected loci is also known as genetic hitch-hiking [94,95,96]. Hitch-hiking mapping aims to detect these regions of low nucleotide diversity within the genome.

Genome wide association studies

Genome-wide association (GWA) mapping makes use of the genome-wide genotyping assays, and is the logical extension of linkage analysis. It is currently a widely applied method for the detection of genetic variation underlying traits in livestock and humans. The aim of GWA studies is to detect statistical association between the phenotype under investigation and assayed markers, with the assumption that the assayed markers are in high LD with the causative variants. The design of a GWA study, especially in complex diseases in human, is typically a case-control design in which healthy individuals are compared to affected individuals. However, for quantitative traits the complete phenotypic distribution for a trait could also be used in the study. In animal genetics, a linear model is typically used to analyze each SNP individually. This linear model usually includes the effect of a SNP, fixed effects, and breeding values of each animal [31].

Unlike linkage analysis, GWA studies do not rely on pedigrees and as a consequence there is no limitation on the number of individuals that can be included in the study. This increased number of individuals will lead to an increased statistical power to detect QTLs with small phenotypic effect and is therefore useful in QTL mapping of traits with a polygenic background [87]. Moreover, as individuals within or between populations will be more distantly related from each other than individuals within a family, more recombination events will have occurred in the original haplotype containing the causative variant. This reduces the shared haplotype that is identical by decent (Box 1) between the individuals, thereby refining the map position of the QTL.

One of the first GWA study in livestock species was published in 2008 using assays containing either 25k or 60k SNPs [88]. The authors identified loci involved in five different monogenic recessive diseases in cattle and for three of them the causative variant was identified after sequencing candidate genes.

Hitch-hiking mapping

Another recently developed method is hitch-hiking mapping [89,90] or selective sweep mapping [91,92,93] (Box 1). Genetic hitch-hiking refers to the process that decreases nucleotide diversity surrounding a selected variant [94,95,96]. Selection on a desirable variant will lead to a reduced or even complete loss of nucleotide diversity at and near the selected locus as non-carrier haplotypes are not selected for in future generations. Hitch-hiking mapping aims to detect regions under selection with the assumption that they must have a functional importance [61]. The challenge of hitch-hiking mapping is to discriminate between regions of selection that are the result of true selection and not from stochastic effect such as

genetic drift [91,92]. Because hitch-hiking mapping is focused on the genetic variation within the genome, it does not require measurable phenotypic information and could, therefore, aid in the detection of genomic regions for traits where phenotypic measurements are difficult, expensive or unethical [90]. Due to the absence of phenotypic information it will, however, not be possible to establish a direct phenotype – genotype relationship [97].

From associated genomic regions to causative variants

Because of the existence of LD, linked markers (assayed marker in LD with the causative variant) can be used for MAS and GS without the need to identify the true causative variant. There are, however, some disadvantages of using these linked markers in breeding. First, linkage between assayed markers and causative variants might not be the same in different populations. This means that for each breeding population, independent studies should be performed to determine the phenotypic effect of assayed markers. Secondly, haplotype decay due to recombination events will decrease LD between linked markers and the causative variant in time. Therefore, new phenotypic based studies are needed every few generation to reexamine the phenotypic effect of the linked markers. The disadvantages for linked markers will be circumvented if direct markers (causative variants) are used in breeding. Direct markers detected in one population will likely have similar phenotypic effects in other populations and recurrent phenotypic studies are not needed. The identification of causative variants is, moreover, essential to understand the biological mechanisms underlying production and disease traits.

The methodology to identify causative variants is straightforward, but nevertheless has proven to be challenging. The first step is to map genomic regions that influence the trait of interest. For large genomic regions fine-mapping might be necessary to reduce the size and number of candidate genes. Fine-mapping could be performed by increasing the marker density within the regions or by including information from additional breeds to identify the minimum haplotype shared identical-by-decent between the breeds [97]. Subsequently genes and other functional elements located within the fine-mapped region are sequenced to detect genetic variants. The identified genetic variants are then subjected to follow-up studies to identify and verify the true functional causative variant. These studies involve screening for clear deleterious variants in genes (such as non-synonymous and frameshift mutations), or will include expression studies or other functional assays.

However, limited successes have been achieved in the detection of causative variants underlying traits in livestock species. Despite the availability of numerous QTL studies, there are currently 12,635 QTLs detected for 1,199 traits in Cattle, Pig, Chicken and Sheep (http://www.animalgenome.org, accessed October 2010), only a limited number of causative variants underlying these QTLs are known [98]. Well known examples of variants underlying QTLs are the SNP in intron 3 of the *IGF2* gene that influences muscle growth in swine [99], and the missense mutation in *DGAT1* that has a major influence on milk yield and composition in cattle [100].
The limited successes to detect causative variants underlying QTLs can be largely contributed to low mapping resolution - most QTL studies are based on linkage analysis- resulting in QTLs mapped to large intervals in the genome. The high costs and labor intensity of traditional Sanger sequencing limited researchers in the detection of all genetic variation located within these QTL regions. To reduce costs, genes with relevant biological function were usually prioritized over genes with unrelated or unknown functions, which could have resulted in neglecting possible important genes. Sequencing efforts were also mainly focused on coding regions, thereby missing possible functional variants in intergenic or regulatory regions. Although the mapping resolution of GWA studies is increased compared to linkage analysis, the identification of underlying causative variants still requires extensive sequencing efforts. However, recent technological developments will provide a rapid and cost effective solution. 'Next generation sequencing' and targeted DNA capture technologies will allow cost effective re-sequencing of entire QTL regions in order to detect underlying genetic variation.

Next generation sequencing technologies

The throughput limitations of classical sequencing using the Sanger enzymatic dideoxy technique [101] initiated efforts to develop new high-throughput sequencing methods for massive parallel sequencing (MPS). Several 'second generation sequencing' technologies are currently available (Roche 454 Life Sciences, Illumina, Life Technologies SOLiD, Helicos Biosciences, Complete Genomics). Each of these technologies is capable of generating millions of short DNA sequences (36-400 bp) in a single run [102,103,104]. The new HiSeq 2000 sequencing platform from Illumina, for instance, is capable of generating up to 350 giga base pairs in one run that takes around 8 days. One run on this platform has the capability to re-sequence the whole genome of nine individual chickens (or three individuals of human, cattle or swine) with a coverage of 30 reads per nucleotide. Developments in MPS technologies are fast and 'third generation sequencing' technologies are currently under development (e.g. Pacific Biosciences,

Ion Torrent, and Oxford Nanopore). These third generation technologies will provide single molecule sequencing, higher throughput capacity, longer reads, and reduced costs and run times.

MPS based strategies are increasingly applied to detect causative variants underlying monogenic and polygenic traits. For instance, MPS has already been used for hitch-hiking mapping in the chicken and for the detection of causative variants involved in monogenic traits in human and polygenic traits in yeast (Box 2).

Box 2. MPS in hitch-hiking mapping, QTL mapping and causative variant detection.

Hitch-hiking mapping

Recently, hitch-hiking mapping using a MPS strategy was performed to detect regions under selection during chicken domestication [106]. In this study, DNA samples of multiple individuals (n=8-11) were pooled to represent the nucleotide diversity within the breeds. These DNA pools were subsequently sequenced by MPS at a low coverage of 4-5 times. In this study, 8 domesticated and 1 non-domesticated breeds were sequenced. In a sliding window approach, several genomic regions were identified where the heterozygosity was substantially deviating from the average of the genome. A putative region under selection at the *TSHR* gene was confirmed by additional analysis in 271 domesticated breeds. The authors, moreover, identified a non-conservative non-synonymous variant in the *TSHR* gene. This variant was identified in the MPS data and was confirmed by Sanger sequencing. The authors suggest that the selection on *TSHR* might be involved in the absence of strict regulation of seasonal reproduction observed in domestic chickens.

Monogenic traits

Whole exome sequencing:

There has, recently, been a rapid increase in the number of papers describing the detection of causative variants involved in monogenic traits in humans [107,108,109,110]. In these studies, the exome of one or more affected and unaffected individuals were re-sequenced to identify causative variants in coding regions of the genome. The exome was isolated from genomic DNA using hybridization based capture methods subsequently followed by MPS. To reduce the number of candidate causative variants, filters were applied to all genetic variants identified. These filters generally included functional importance (non-synonymous, splice acceptor and donor site variants, or indels), absence in non-affected individuals, and allele frequencies in the population (for instance, common variants are likely not to be involved in rare diseases). As an example, Ng [109] identified the causative gene for Miller syndrome, a rare recessive syndrome in humans. Exome sequencing of 4 affected individuals resulted in 1,525 candidate genes in which two or more functional important variants were detected. In the study eight non-affected individuals were also re-sequenced with MPS. After applying a filter for common SNPs, and absence or presence in the non-affected individuals, only one gene - *DHODH* - remained as a candidate gene. Additional sequencing efforts in affected individuals resulted in the identification of 12 causative variants within this gene.

Box 2. Continued...

Whole genome re-sequencing (family based):
Roach *et al*. [111] re-sequenced the whole genome of a family of four, consisting of two affected offspring and their non-affected parents. Both affected offspring were affected by the homozygous diseases Miller syndrome and primary ciliary dyskinesia, of which causative genes had already been identified. The family structure enabled inheritance analysis, which led to the identification of sequencing errors and precise locations of recombination events. Because both offspring were affected with a recessive trait, follow-up studies were limited to 22% of the genome in which both offspring were identical. The recessive inheritance of the disease, moreover, requires that both non-affected parents are heterozygous for the causative variant (either heterozygous for the same causative variant, or compound heterozygous variants within the same gene). Focusing only on rare variants, four candidate genes remained, including the previously known causative genes for Miller syndrome and primary ciliary dyskinesia.

Whole genome re-sequencing combined with linkage analysis:
Sobreira *et al.* [112] combined MPS with classical linkage analysis to reduce the number of individuals that needed to be fully sequenced. Linkage analysis using genome-wide SNPs assays in a family of 12 members (7 affected) resulted in the association of six genomic regions (42 Mb in total) associated with the autosomal dominant trait metachondromatosis. Whole genome sequencing (32X coverage) of a single affected individual within this family provided the identification of a single deleterious variant - an 11bp deletion in *PTNP11* resulting in a frameshift and premature stop codon - within one of the six associated regions. Subsequent sequencing of *PTNP11* in the other affected individuals resulted in the conformation of this variant. Subsequently, sequencing the *PTNP11* gene in a second family with metachondromatosis resulted in the identification of nonsynonymous variant that also resulted in a premature stop. Without the linkage analysis, follow-up studies to determine the causative gene and underlying variants would have been much more complicated. Instead of the single candidate variant identified in the 42 Mb of associated regions, the authors would have needed follow-up analysis on each of the 109 protein-truncating variants identified within the genome of this individual.

Polygenic traits
Extreme QTL mapping or extreme-trait re-sequencing involves the selection of a small number of individuals or DNA pools with extreme phenotypes - for instance disease resistant or estimated breeding values - of the trait of interest [113]. Ehrenreich *et al.* [114] applied this extreme QTL mapping in order to identify loci involved in sensitivity to 4-nitroquiloline (4-NQO), a polygenic trait in yeast. Two previously identified loci only explained a small part of the genetic variation of sensitivity to 4-NQO, implying that additional loci must exist for this trait. In the study, DNA pools of two yeast cultures with different phenotypes were re-sequenced with MPS. The 'resistant' culture was grown on medium containing 4-NQO and therefore contained only yeast that were resistant to 4-NQO. The 'control' group was grown on normal medium and therefore contained yeast both susceptible and resistant to 4-NQO. For both cultures, DNA was extracted and subsequently sequenced by MPS. Comparison between the two cultures resulted in the detection of 14 loci where the allele frequency of the resistant culture was significantly different from the control culture. The 14 loci,

including the two previously identified loci, explained 70% of the total genetic variance of the trait.

1.2 Aim and outline of this thesis

The research described in this thesis aimed to investigate the utility of several molecular approaches to (i) to identify causative variants underlying monogenic (chapter 3) or polygenic traits (chapter 4), (ii) map genomic regions that are or have been under selection in the chicken genome (chapter 5), and (iii) to improve and increase available genomic resources in the chicken (chapter 2 and 4).

In **chapter 2** of this thesis, we describe the construction of a new high resolution linkage map of the chicken genome based on 1,617 animals of two broiler population of which each individual was genotyped with a genome-wide assay containing 17,790 SNPs. The main goals of this linkage map were to assist in the improvement of the current genome assembly by mapping 613 previously unmapped markers and to provide a high resolution linkage map for linkage analysis and association studies. The high resolution linkage map generated in this chapter, moreover, allowed us to discuss on recombination rates across the genome between the two mapping populations. In **chapter 3**, we describe the molecular characterization of the locus causing the late feathering phenotype; a monogenic trait in chicken that results in a delayed emergence of flight feathers at hatch. The late feathering phenotype is beneficial to breeders as it can be used for sex typing at hatch. The locus has, therefore, been extensively used in diverse commercial chicken breeds. However, a retrovirus closely linked to the late feathering allele causes a negative pleiotropic effect of this locus on egg production and viral infections. The identification of the causative variant underlying the late feathering phenotype will allow screening for recombinants between the beneficial late feathering allele and the undesired retrovirus. Within this chapter we describe the identification of a 180 kb tandem duplication in the late feathering allele using a quantitative PCR approach. We, moreover, describe a molecular test to specifically detect this duplication, also in heterozygous individuals. In **chapter 4**, we combined a GWA study with MPS to detect causative variants underlying the pulmonary hypertension syndrome (PHS) in chicken. PHS is a polygenic trait that causes substantial financial losses in the poultry industry and results in reduced animal welfare. In this study we performed a GWA study to detect QTLs associated with PHS. For variant detection, we used MPS to sequence the genomes of twelve broiler chickens. To maximize the occurrence of variants involved in PHS, we selected 6 animals with an extreme low estimated breeding value, and 6 animals with an extreme high estimated breeding value. Within this study we focused on protein affecting variants located within the QTL regions. In addition, this chapter

describes the identification of 7.62 million SNPs that will aid in improving future genome-wide assays. We describe a hitch-hiking mapping method in **chapter 5** to detect signatures of selection in the genome of 67 commercial and non-commercial chicken breeds. For this mapping strategy we genotyped pooled DNA samples from each breed using a genome-wide assay including nearly 60,000 SNPs. In this chapter we discuss on several regions under selection that were identified, and we discuss on underlying candidate genes that might be involved in production or disease traits. Finally, in the general discussion described in **chapter 6**, I discuss on the main findings of this thesis and comment on the strategies to identify causative variants underlying production and disease traits.

Definitions

Variant:
Any allele that is inherited by Mendelian laws, and includes single nucleotide polymorphisms (SNP), indels, structural variations and irreversible epigenetic modifications.

Structural variation:
Copy number variations (CNV, large (>1kb) insertions, deletions, and duplications) and copy neutral variation (inversions and translocations).

Causative variant:
A variant that is the true underlying cause of a phenotypic effect.

Trait:
Any phenotypic observation that can be made within an organism. In this thesis a trait mainly refers to production traits, fitness traits, and disease resistance or susceptibility.

Selection:
Directional selection, either natural or artificial, that results in either an increased or decreased frequency of the variant under selection.

References

1. Hedges SB (2002) The origin and evolution of model organisms. Nature Rev Genet 3: 838-849.
2. Crawford RD (1990) Poultry Breeding and Genetics. Elsevier Science, New York.
3. Darwin C (1868) The Variation of Animals and Plants under Domestication. Macmillan Publishers Limited.
4. Danforth C (1958) Gallus sonnerati and the domestic fowl. J Hered 49: 167-169.
5. Steiner H (1945) Ueber letal Fehentwicklung der Zeiten Zackommenschafts-Generation bei tieischen Arbastarden Arcb. Arcb Julius Klaus-Stift 20: 236-251.
6. Baker C (1968) Molecular genetics of avian proteins. IX. Interspecific and intraspecific variation of egg white proteins of the genus gallus. Genetics 58: 211-226.
7. Frisby DP, Weiss RA, Roussel M, Stehelin D (1979) The distribution of endogenous chicken retrovirus sequences in the DNA of galliform birds does not coincide with avian phylogenetic relationships. Cell 17: 623-634.
8. Fumihito A (1994) One subspecies of the red jungle fowl (Gallus gallus gallus) suffices as the matriarchic ancestor of all domestic breeds. Proc Natl Acad Sci USA 91: 12505-12509.
9. Fumihito A, Miyake T, Takada M, Shingu R, Endo T, *et al.* (1996) Monophyletic origin and unique dispersal patterns of domestic fowls. Proceedings of the National Academy of Sciences of the United States of America 93: 6792-6795.
10. Hillel J, Groenen MA, Tixier-Boichard M, Korol AB, David L, et al (2003) Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. Genet Sel Evol 35: 533-557.
11. Hutt FB (1949) Genetics of the Fowl. McGraw Hill Book Company Inc., New York.
12. Plant J (1986) The origin, evolution, history and distribution of the domestic fowl. Part 3. The Gallus species. Jungle fowls. 5 Bonar street, Maitland 2320, N.S.W., Australia: Privately published.
13. Eriksson J (2008) Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. PLoS Genet 4: e1000010.

14. Liu Y, Wu G, Yao Y, Miao Y, Luikart G, *et al.* (2006) Multiple maternal origins of chickens: out of the Asian jungles. Mol Phylogenet Evol 38: 12 - 19.
15. Oka T, Ino Y, Nomura K, Kawashima S, Kuwayama T, *et al.* (2007) Analysis of mtDNA sequences shows Japanese native chickens have multiple origins. Animal Genetics 38: 287-293.
16. Kanginakudru S, Metta M, Jakati R, Nagaraju J (2008) Genetic evidence from Indian red jungle fowl corroborates multiple domestication of modern day chicken. BMC Evolutionary Biology 8: 174.
17. West B, Zhou B-X (1988) Did chicken go north? New evidence for domestication. J Archaeol Sci 15: 515 - 533.
18. Skinner J (1974) American poultry history, 1823-1973 America Printing and Publishing, Inc, Madison, Wisconsin.
19. Havenstein GB, Ferket PR, Qureshi MA (2003) Carcass composition and yield of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. Poult Sci 82: 1509-1518.
20. Burt DW (2005) Chicken genome: Current status and future opportunities. Genome Research 15: 1692-1698.
21. Baghbanzadeh A, Decuypere E (2008) Ascites syndrome in broilers: physiological and nutritional perspectives. Avian Pathology 37: 117 - 126.
22. Muir W, Wong G, Zhang Y, Wang J, Groenen M, *et al.* (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proc Natl Acad Sci USA 105: 17312 - 17317.
23. Decuypere E, Bruggeman V, Barbato G, Buyse J (2003) Problems associated with selection for increased broiler meat production: Growth and reproduction. In: M. Muir, S. E. Aggrey, and G. W. Keeton, editor. Poultry Genetics, Breeding and Biotechnology: CABI Publishing, CAB International. pp. 13-28.
24. Zekarias B, Huurne AAHMT, Landman WJM, Rebel JMJ, Pol JMA, *et al.* (2002) Immunological basis of differences in disease resistance in the chicken. Vet Res 33: 109-125.
25. Julian RJ (1998) Rapid growth problems: ascites and skeletal deformities in broilers. Poult Sci 77: 1773-1780.
26. Whitehead C, Fleming R (2000) Osteoporosis in cage layers. Poult Sci 79: 1033-1041.
27. Julian RJ (1993) Ascites in poultry. Avian Pathology 22: 419 - 454.

28. Balog JM (2003) Ascites Syndrome (Pulmonary Hypertension Syndrome) in Broiler Chickens: Are We Seeing the Light at the End of the Tunnel? Avian and Poultry Biology Reviews 14: 99 -126.
29. Rabie T, Crooijmans R, Bovenhuis H, Vereijken A, Veenendaal T, *et al.* (2005) Genetic mapping of quantitative trait loci affecting susceptibility in chicken to develop pulmonary hypertension syndrome. Animal Genetics 36: 468 - 476.
30. Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, *et al.* (2010) Food Security: The Challenge of Feeding 9 Billion People. Science 327: 812-818.
31. Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet.: 10: 381-391.
32. Neimann-Sorensen A, Robertson A (1961) The association between blood groups and several production characteristics in three Danish cattle breeds. Acta Agriculturae Scandinavica 11: 163-196.
33. Smith C (1967) Improvement of metric traits through specific genetic loci. Animal Science 9: 349-358.
34. Lande R, Thompson R (1990) Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. Genetics 124: 743-756.
35. Meuwissen T, Goddard M (1996) The use of marker haplotypes in animal breeding schemes. Genetics Selection Evolution 28: 161 - 176.
36. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157: 1819-1829.
37. Schmutz J, Grimwood J (2004) Genomes: Fowl sequence. 432: 679-680.
38. Brown W, Hubbard S, Tickle C, Wilson S (2003) The chicken as a model for large-scale analysis of vertebrate gene function. Nat Rev Genet 4: 87 - 98.
39. Bumstead N, Palyga J (1992) A preliminary linkage map of the chicken genome. Genomics 13: 690-697.
40. Crittenden L, Provencher L, Levin I, Abplanalp H, Briles R, *et al.* (1993) Characterization of a red jungle fowl by white leghorn backcross reference population for molecular mapping of the chicken genome. Poultry Science 72: 334-348.
41. Jacobsson L, Park H, Wahlberg P, Jiang S, Siegel P, *et al.* (2004) Assignment of fourteen microsatellite markers to the chicken linkage map. Poult Sci 83: 1825-1831.

42. Kerje S (2003) The twofold difference in adult size between the red junglefowl and white leghorn chickens is largely explained by a limited number of QTLs. Anim Genet 34: 264-274.
43. Groenen M, Cheng H, Bumstead N, Benkel B, Briles W, *et al.* (2000) A consensus linkage map of the chicken genome. Genome Res 10: 137 - 147.
44. Groenen M, Wahlberg P, Foglio M, Cheng H, Megens H, *et al.* (2009) A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res 19: 510 - 519.
45. Elferink M, van As P, Veenendaal T, Crooijmans R, Groenen M (2010) Regional differences in recombination hotspots between two chicken populations. BMC Genetics 11: 11.
46. Morisson M, Lemière A, Bosc S, Galan M, Plisson-Petit F, *et al.* (2002) ChickRH6: a chicken whole-genome radiation hybrid panel. Genet Sel Evol 34: 521-533.
47. Jennen DGJ, Crooijmans RPMA, Morisson M, Grootemaat AE, van der Der Poel JJ, *et al.* (2004) A radiation hybrid map of chicken chromosome 15. Animal Genetics 35: 63-65.
48. Rabie TSKM, Crooijmans RPMA, Morisson M, Andryszkiewicz J, van der Poel JJ, *et al.* (2004) A radiation hybrid map of chicken Chromosome 4. Mammalian Genome 15: 560-569.
49. Boardman P, Sanz-Ezquerro J, Overton I, Burt D, Bosch E, *et al.* (2002) A comprehensive collection of chicken cDNAs. Curr Biol 12: 1965 - 1969.
50. Hubbard SJ, Grafham DV, Beattie KJ, Overton IM, McLaren SR, *et al.* (2005) Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags. Genome Research 15: 174-183.
51. Crooijmans R, Vrebalov J, Dijkhof R, van der Poel J, Groenen M (2000) Two-dimensional screening of the Wageningen chicken BAC library. Mamm Genome 11: 360 - 363.
52. Lee M, Ren C, Yan B, Cox B, Zhang H, *et al.* (2003) Construction and characterization of three BAC libraries for analysis of the chicken genome. Anim Genet 34: 151-152.
53. Ren C, Lee M-K, Yan B, Ding K, Cox B, *et al.* (2003) A BAC-Based Physical Map of the Chicken Genome. Genome Research 13: 2754-2758.

54. Aerts J, Crooijmans R, Cornelissen S, Hemmatian K, Veenendaal T, *et al.* (2003) Integration of chicken genomic resources to enable whole-genome sequencing. Cytogenetic and Genome Research 102: 297-303.
55. Wallis J, Aerts J, Groenen M, Crooijmans R, Layman D, *et al.* (2004) A physical map of the chicken genome. Nature 432: 761 - 764.
56. Cheng H, Levin I, Vallejo R, H K, Dodgson J, *et al.* (1995) Development of a genetic map of the chicken with markers of high utility. Poult Sci 74: 1855 - 1874.
57. Crooijmans R, van Oers P, Strijk J, van der Poel J, Groenen M (1996) Preliminary linkage map of the chicken (Gallus domesticus) genome based on microsatellite markers: 77 new markers mapped. Poultry Science 75: 746-754.
58. Wong G, Liu B, Wang J, Zhang Y, Yang X, *et al.* (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432: 717 - 722.
59. Hillier L, Miller W, Birney E, Warren W, Hardison R, *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695 - 716.
60. Kelley D, Salzberg S (2010) Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biology 11: R28.
61. Nielsen R (2005) Molecular signatures of natural selection. Annual Review of Genetics 39: 197-218.
62. Tobita-Teramoto T, Jang G, Kino K, Salter D, Brumbaugh J, *et al.* (2000) Autosomal albino chicken mutation (ca/ca) deletes hexanucleotide (-deltaGACTGG817) at a copper-binding site of the tyrosinase gene. Poult Sci 79: 46-50.
63. Kerje S, Lind J, Schütz K, Jensen P, Andersson L (2003) Melanocortin 1-receptor (MC1R) mutations are associated with plumage colour in chicken. Animal Genetics 34: 241-248.
64. Kerje S, Sharma P, Gunnarsson U, Kim H, Bagchi S, *et al.* (2004) The Dominant white, Dun and Smoky Color Variants in Chicken Are Associated With Insertion/Deletion Polymorphisms in the PMEL17 Gene. Genetics 168: 1507-1518.
65. Honkatukia M, Reese K, Preisinger R, Tuiskula-Haavisto M, Weigend S, *et al.* (2005) Fishy taint in chicken eggs is associated with a substitution within a conserved motif of the FMO3 gene. Genomics 86: 225-232.

66. Klungland H, Våge D, Gomez-Raya L, Adalsteinsson S, Lien S (1995) The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. Mamm Genome September: 636-639.
67. Seitz JJ, Schmutz SM, Thue TD, Buchanan FC (1999) A missense mutation in the bovine MGF gene is associated with the roan phenotype in Belgian Blue and Shorthorn cattle. Mammalian Genome 10: 710-712.
68. Mariat D, Taourit S, Guérin G (2003) A mutation in the MATP gene causes the cream coat colour in the horse. Genet Sel Evol 35: 119-133.
69. Marklund L, Moller M, Sandberg K, Andersson L (1996) A missense mutation in the gene for melanocyte-stimulating hormone receptor (MCIR) is associated with the chestnut coat color in horses. Mammalian Genome 7: 895-899.
70. Kijas JMH, Wales R, Tornsten A, Chardon P, Moller M, *et al.* (1998) Melanocortin Receptor 1 (MC1R) Mutations and Coat Color in Pigs. Genetics 150: 1177-1185.
71. Grobet L, Royo Martin LJ, Poncelet D, Pirottin D, Brouwers B, *et al.* (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. Nat Genet 17: 71-74.
72. Charlier C, Segers K, Karim L, Shay T, Gyapay G, *et al.* (2001) The callipyge mutation enhances the expression of coregulated imprinted genes in cis without affecting their imprinting status. Nat Genet 27: 367-369.
73. Fujii J, Otsu K, Zorzato F, de Leon S, Khanna V, *et al.* (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. Science 253: 448-451.
74. Milan D, Jeon J-T, Looft C, Amarger V, Robic A, *et al.* (2000) A Mutation in PRKAG3 Associated with Excess Glycogen Content in Pig Skeletal Muscle. Science 288: 1248-1251.
75. Lin L, Faraco J, Li R, Kadotani H, Rogers W, *et al.* (1999) The Sleep Disorder Canine Narcolepsy Is Caused by a Mutation in the Hypocretin (Orexin) Receptor 2 Gene. Cell 98: 365-376.
76. Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome.Nat. Rev. Genet.: 3: 299-309.
77. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-1517.
78. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95-108.

79. Boehnke M (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Hum Genet 55(2): 379–390.
80. Andersson L (2001) Genetic dissection of phenotypic diversity in farm animals. Nature Rev Genet 2: 130-138.
81. Muir W, Wong G, Zhang Y, Wang J, Groenen M, *et al.* (2008) Review of the Initial Validation and Characterization of a 3 K Chicken SNP Array. World's Poultry Science Journal 64: 219 - 226.
82. Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. Harlow, Essex, UK.: Longmans Green.
83. Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, *et al.* (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29: 233-237.
84. Rodionov AV (1996) Micro vs. macro: structural-functional organization of avian micro- and macrochromosomes. Genetika 32: 597-608.
85. Megens H-J, Crooijmans R, Bastiaansen J, Kerstens H, Coster A, *et al.* (2009) Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genetics 10: 86.
86. Aerts J, Megens H, Veenendaal T, Ovcharenko I, Crooijmans R, *et al.* (2007) Extent of linkage disequilibrium in chicken. Cytogenet Genome Res 117: 338 - 345.
87. Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405: 847-856.
88. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, *et al.* (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. 40: 449-454.
89. Harr B, Kauer M, Schlötterer C (2002) Hitchhiking mapping: A population-based fine-mapping strategy for adaptive mutations in Drosophilamelanogaster. Proceedings of the National Academy of Sciences of the United States of America 99: 12949-12954.
90. Schlotterer C (2004) The evolution of molecular markers--just a matter of fashion? Nat Rev Genet 5: 63 - 69.
91. Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD (2005) Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. Genetics 170: 1401-1410.
92. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, *et al.* (2005) Genomic scans for selective sweeps using SNP data. Genome Research 15: 1566-1575.

93. Pavlidis P, Hutter S, Stephan W (2008) A population genomic approach to map recent positive selection in model species. Molecular Ecology 17: 3585-3598.
94. Kojima K-i, Schaffer HE (1967) Survival Process of Linked Mutant Genes. Evolution 21: 518-531.
95. Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genetics Research 23: 23-35.
96. Berry AJ, Ajioka JW, Kreitman M (1991) Lack of Polymorphism on the Drosophila Fourth Chromosome Resulting From Selection. Genetics 129: 1111-1117.
97. Andersson L, Georges M (2004) Domestic-animal genomics: deciphering the genetics of complex traits. Nature Rev Genet 5: 202-212.
98. Ron M, Weller JI (2007) From QTL to QTN identification in livestock – winning by points rather than knock-out: a review. Animal Genetics 38: 429-439.
99. Van Laere A-S, Nguyen M, Braunschweig M, Nezer C, Collette C, *et al.* (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425: 832-836.
100. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, *et al.* (2002) Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. Genome Research 12: 222-231.
101. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74: 5463–5467.
102. Mardis ER (2008) Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics 9: 387-402.
103. Pushkarev D, Neff NF, Quake SR (2009) Single-molecule sequencing of an individual human genome. Nat Biotech 27: 847-850.
104. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, *et al.* (2010) Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 327: 78-81.
105. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, *et al.* (2002) Genetic Structure of Human Populations. Science 298: 2381-2385.
106. Rubin C, Zody MC, Eriksson J, Meadows JRS, Sherwood E, *et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464: 587-591.

107. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. 461: 272-276.
108. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, *et al.* (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. 42: 790-793.
109. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. 42: 30-35.
110. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences 106: 19096-19101.
111. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, *et al.* (2010) Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science 328: 636-639.
112. Sobreira NLM, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, *et al.* (2010) Whole-Genome Sequencing of a Single Proband Together with Linkage Analysis Identifies a Mendelian Disease Gene. PLoS Genet 6: e1000991.
113. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet.: 11: 415-425.
114. Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, *et al.* (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature: 464: 1039-1042.

2

Regional differences in recombination hotspots between two chicken populations

M.G. Elferink[1], P. van As[2], A. Veenendaal[1], R.P.M.A. Crooijmans[1], M.A.M. Groenen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University and Research Centre, Marijkeweg 40, PO Box 338, Wageningen, The Netherlands;
[2] Hendrix Genetics Research, Technology & Services B.V., Spoorstraat 69, PO Box 114, Boxmeer, the Netherlands

BMC Genetics (2008) 11:11

Abstract

Although several genetic linkage maps of the chicken genome have been published, the resolution of these maps is limited and does not allow the precise identification of recombination hotspots. The availability of more than 3.2 million SNPs in the chicken genome and the recent advances in high throughput genotyping techniques enabled us to increase marker density for the construction of a high-resolution linkage map of the chicken genome. This high-resolution linkage map allowed us to study recombination hotspots across the genome between two chicken populations: a purebred broiler line and a broiler x broiler cross. In total, 1,619 animals from the two different broiler populations were genotyped with 17,790 SNPs.

The resulting linkage map comprises 13,340 SNPs. Although 360 polymorphic SNPs that had not been assigned to a known chromosome on chicken genome build WASHUC2 were included in this study, no new linkage groups were found. The resulting linkage map is composed of 31 linkage groups, with a total length of 3,054 cM for the sex-average map of the combined population. The sex-average linkage map of the purebred broiler line is 686 cM smaller than the linkage map of the broiler x broiler cross.

In this study, we present a linkage map of the chicken genome at a substantially higher resolution than previously published linkage maps. Regional differences in recombination hotspots between the two mapping populations were observed in several chromosomes near the telomere of the p arm; the sex-specific analysis revealed that these regional differences were mainly caused by female-specific recombination hotspots in the broiler x broiler cross.

2.1 Introduction

Genetic linkage maps are essential to identify genomic regions that influence complex phenotypes (quantitative trait loci), to assist in the sequence assembly of genomes, and to study recombination across the genome. Linkage analysis and genome-wide association studies not only require high marker densities, but also accurate linkage maps in order to detect quantitative trait loci [1]. High-density linkage maps have been described for humans [2-4], mice [5], rats [6], and chickens [7]. Chicken linkage maps have been published ranging from 100 RFLP markers [8] to a high-density map comprising thousands of markers, most of which are single nucleotide polymorphisms (SNP) [7].

In combination with a physical BAC contig map [9], linkage maps of the chicken [10-12] were used to construct the draft genome sequence of the chicken. The draft sequence of the chicken genome, published in 2004, comprises 1.05 Gb [13]. In chicken genome build WASHUC2 (May 2006) there were a total of 997 Mb of assigned sequences, which covered the two sex chromosomes (Z and W) and 29 of the 38 autosomes. The unassembled sequences that remained were combined in chromosome unassigned. The most recent linkage map, published in 2009 by Groenen *et al.*, consists of 34 different linkage groups (including GGAZ); thus, at least five autosomal chromosomes are still entirely unrepresented [7].

Differences in the sizes of the linkage map were found among several chicken populations [7, 10, 11, 14, 15]. In these studies, domesticated populations tended to have increased recombination compared to nondomesticated populations. This finding was in agreement with the hypothesis that selection leads to higher rates of recombination [16]. Due to the limited resolution of the published chicken linkage maps the specific underlying regions where recombination differs among the chicken populations could not be identified. Moreover, these studies mainly focused on sex-average recombination, and did not take into account the influence of sex on recombination in chickens.

The availability of more than 3.2 million SNPs in the chicken genome (dbSNP build 128 and [17]) and the recent advances in high-throughput genotyping techniques makes it feasible to increase marker density for linkage analysis and genome-wide association studies and to study recombination rates across the genome in the chicken.

In this study, we present a high-resolution linkage map of the chicken genome based on data from a cross between two different broiler lines (n= 306) and on data from a different single purebred broiler line (n= 1313). Both populations were genotyped with an 18K SNP Illumina Infinium iSelect Beadchip. The high-resolution

linkage maps generated in this study allowed us to study regions of recombination hotspots between the two mapping populations and between the sexes.

2.2 Methods

Marker selection

In total, 17,790 markers were included on the Illumina Infinium iSelect Beadchip (Additional File 1). Markers were selected from dbSNP build 122. The Beadchip consisted of 17,177 markers that had been mapped and 613 markers that had not been mapped to a chromosome or linkage group. Markers were distributed evenly across each chromosome, with marker densities based on the size of the chromosome. For GGA1–GGA5 and GGAZ, markers were selected every 50 kb; for GGA6–GGA10 every 36 kb; for GGA11–GGA20 every 25 kb; and for GGA21–GGA28 every 15.5 kb. Two additional linkage groups, which were not assigned to a chromosome, were also included on the beadchip: LGE22C19W28_E50C23 (from here on called LGE22) and LGE64. The unmapped markers were located on contigs larger than 100,000 bp, which were found in the unassigned sequences of the draft sequence (chromosome unassigned). The 613 markers were selected randomly, except for the size of the contig in which they were located.

Genotyping was performed using the standard protocol for Infinium iSelect Beadchips. Data were analyzed with Beadstudio Genotyping v3.0.19.0, and quality control was performed according to the guidelines from the Infinium genotyping data analysis protocol [18].

Populations

In total, 1,619 animals from two populations were genotyped with the 18K SNP beadchip. Blood and DNA sample collection was carried out by licensed and authorized personnel under approval of Hendrix Genetics. Population 1 was an advanced intercross line derived from a cross between two broiler dam lines [19, 20]. The maternal line was selected for reproduction (egg numbers as the most important trait, as well as hatching of fertile eggs) and, to a lesser extent, body weight. The maternal line was not selected for feed conversion rate and breast meat percentage. The paternal line was selected for growth and feed conversion rate (almost equally important), and selection with regard to reproduction was performed to keep performance constant (it also compensated the negative effects of selection for growth). The paternal line, moreover, was also subject to some selection for conformation. There was no selection for breast meat percentage for this line. The maternal and paternal lines both originated from the White Plymouth Rock breed. Population 1 was used previously for quantitative trait loci mapping of

pulmonary hypertension syndrome [19, 21]; fatness traits in broilers [22]; and bodyweight, growth rate and feed efficiency [23, 24]. Combined with other populations, a subset of population 1 has previously been used to construct the consensus linkage map of the chicken genome [7, 12]. In total, 306 animals were genotyped from population 1: 10 full-sib families of generation 1; 20 parents (10 males and 10 females) and 50 offspring (11 males and 39 females); and 37 full- and half-sib families of generation 6 or 7; 66 parents (32 males and 34 females) and 170 offspring (61 males, 67 females, and 42 of unknown sex). Population 2 consisted of a third purebred commercial broiler dam line that was selected for breast meat percentage. This population also originated from the White Plymouth Rock breed. In total, 1,313 animals were genotyped from population 2: 266 parents (68 males and 198 females) and 1,047 offspring (107 males and 940 females).

Linkage analysis

The linkage map was constructed with a modified version of CRI-MAP [25]. This modified version can handle large datasets and was provided by Drs. Liu and Grosz of Monsanto Company (St. Louis, MO, USA). During construction of the linkage map, a marker was considered to be informative if it had at least 20 informative meiosis. The linkage map was constructed with the use of five options: AUTOGROUP, BUILD, CHROMPIC, FLIPSN, and FIXED. AUTOGROUP was used to check each chromosome unassigned marker for linkage to a known chromosomes or linkage groups (thresholds used: LOD = 4, informative meiosis = 0, different chromosomes = 5, and linkage ratio = 0.5). Markers were assigned to a specific chromosome if linkage was found, or remained in chromosome unassigned if no linkage was found. The initial marker order was similar to the order in which the markers were located on the physical map (WASHUC2 build, May 2006). The BUILD option was used to determine the most likely position of the newly assigned markers in the marker order. Markers were mapped to a specific position if BUILD incorporated the marker at one specific position only (threshold LOD = 3). If multiple positions were found, the best position was based on three criteria: (1) if the sequence of the contig in which the marker was located showed a (partial) BLAST hit against one of the possible locations indicated by BUILD, (2) if one of the positions in the BUILD output had a higher LOD score (>1) than all other positions and, (3) if a gap was found between two (super) contigs on the physical map. If no specific position was found using these criteria, the marker was excluded from the analysis. The BUILD output was, furthermore, used to determine potential errors in the marker order. Markers that showed high recombination rates compared to flanking markers (>3cM on both sides) were taken out of the map and reanalyzed

by BUILD. CHROMPIC was used to identify double recombinants, which, at the marker density used, are a good indication of marker order errors or genotype errors. Double-recombinant markers were reanalyzed by BUILD to determine the most likely position. Double recombinants that could not be resolved after repositioning were most likely caused by genotyping errors, and were therefore removed from the dataset. FLIPSN (*n* = 5) was also used to correct errors in the marker order. If an alternative marker order was more likely than the initial one (LOD increased by >1), the new marker order was used. To decrease errors and increase the accuracy of the map, the CHROMPIC, BUILD, and FLIPSN options were used repeatedly for each chromosome until no double recombinants were observed and the most likely marker order was achieved for the remaining markers. Finally, the FIXED option was used to construct the sex-specific and sex-average linkage maps. For the markers that remained in the chromosome unassigned, TWOPOINT analyses were performed to find linkage between the markers (LOD = 3).

Recombination rate

Recombination rates were calculated for nonoverlapping bins of approximately 500kb. Linkage maps for population 1 and 2 were constructed with all of the markers that were informative in at least one of the populations. The recombination rate of each bin is expressed as the genetic length in centimorgans divided by the genomic length in mega base pairs.

Statistical Analysis

To test if differences in map distances between populations differed significantly we assumed that 1 cM equals a recombination fraction of 0.01 and calculated the Z-test statistic as

$$Z = \frac{\theta_1 - \theta_2}{\sqrt{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}}}$$

where

θ_1 = the recombination fraction in population 1,

θ_2 = the recombination fraction in population 2,

n_1 = the average number of informative meiosis in population 1,

n_2 = the average number of informative meiosis in population 2.

p–values were obtained from a standard normal distribution. Recombination fractions were determined for sliding windows consisting of eight bins. When

differences in recombination fractions between males and females were tested, it was assumed that both sexes contributed equally to the number of informative meiosis. We considered a nominal p<0.01 as suggestive evidence for differences in recombination fraction. Further, for results to be significant, a more stringent significance criteria $p < 2.46*10^{-4}$ was defined that accounts for multiple testing along the genome. Multiple testing was accounted for by applying a Bonferroni correction assuming 203 independent tests and a nominal α=0.05. For in total 1624 "windows" differences in recombination were determined, however, as a result of the sliding window approach (a window consisting of eight bins), every 8^{th} sliding window is truly independent which results in 203 independent tests.

2.3 Results

Linkage analysis

In total, 13,340 informative markers (75% of all markers on the SNP beadchip) and 1,619 individuals were used to construct the combined linkage map of the two populations (Additional Files 2 and 3). In total, 613 markers that had not been mapped to a known chromosome or linkage group were included on the beadchip. Of the 613 unassigned markers, 103 did not pass quality control, 150 were homozygous, and 360 were informative (Additional Files 3 and 4). Of the informative markers, 343 could be assigned to a known chromosome or linkage group, and 17 could not. These 17 markers also showed no linkage to each other, even when the LOD score threshold was set to 2. From the 343 markers that were assigned to a known chromosome with AUTOGROUP, 230 were included in the final linkage map. No specific position on a chromosome could be determined for the remaining 110 markers (three GGW assigned markers were not included in the analysis), and they were therefore not included in the linkage map.

As a starting point for building the linkage map, we used the marker order based on the position of the markers on the sequence map. In general, this order appeared to be in agreement with the most likely marker order for the linkage map. Some adjustments, nevertheless, were made: on GGA5, a block of thirteen markers was inverted, which resulted in a 1.4 cM decrease in the size of the map; on GGA13, five markers were inverted, which resulted in a 1.5 cM decrease in the size of the map; in linkage group LGE22, rearranged markers resulted in a decrease of 3.1 cM; and in linkage group LGE64, rearranged markers resulted in a decrease of 9.4 cM.

Table 2.1 The linkage map lengths and recombination rates for the chicken chromosomes of the combined populations.

Chromosome	Length[1]	Sex-average	Sex-specific		Recombination rate
			Female	Male	
	(Mb)	(cM)	(cM)	(cM)	(cM/Mb)
GGA1	200.9	413.5	377.1	455.3	2.1
GGA2	154.8	281.3	259.9	303.5	1.8
GGA3	113.6	236.9	225.6	250.2	2.1
GGA4	94.2	195.2	182.5	207.7	2.1
GGA5	62.2	154.4	154.9	155.1	2.5
GGA6	37.3	93.8	85.0	102.4	2.5
GGA7	38.3	103.1	99.0	107.3	2.7
GGA8	30.6	96.6	94.2	98.9	3.2
GGA9	25.5	88.1	85.4	91.1	3.5
GGA10	22.5	80.6	79.6	81.1	3.6
GGA11	21.9	64.0	63.3	64.9	2.9
GGA12	20.5	69.1	67.9	70.7	3.4
GGA13	18.9	62.7	63.8	61.6	3.3
GGA14	15.8	67.4	72.5	65.2	4.3
GGA15	13.0	53.6	52.9	54.2	4.1
GGA16	0.4	55.6	59.1	53.5	n.d.[2]
GGA17	11.2	50.9	51.5	51.0	4.6
GGA18	10.9	51.7	49.9	53.5	4.7
GGA19	9.9	52.3	53.2	52.0	5.3
GGA20	13.9	55.1	55.2	54.8	4.0
GGA21	6.9	56.9	57.2	56.5	8.2
GGA22	3.9	56.4	59.9	52.4	14.3
GGA23	6.0	52.3	51.4	53.0	8.7
GGA24	6.4	53.2	53.4	52.4	8.3
GGA25	2.0	57.1	54.0	59.4	n.d.[2]
GGA26	5.1	52.3	51.4	52.9	10.3
GGA27	4.7	51.0	50.6	51.5	10.8
GGA28	4.5	53.6	52.5	54.3	11.9
LGE22	0.9	59.3	58.5	64.5	n.d.[2]
LGE64	0.0	8.4	6.7	8.7	n.d.[2]
GGAZ	74.6	227.7	-	227.1	3.0
Total autosomal	956.9	2826.4	2728.1	2939.6	3.0
Total length	1031.5	3053.5	2728.0	3166.7	3.0

[1] Physical length of the chromosome was based on the position of the last marker in the WASHUC2 build.
[2] n.d.= not determined, as the chromosome showed clear evidence of sequence gaps.

The number of informative meiosis per mapped marker for the combined linkage map ranged from 20 to 1,242, with an average of 517. The total length of the sex-average map was 3,053.5 cM (Table 2.1). The female sex-specific map was 211.5

cM smaller than the male sex-specific map, with a female-to-male ratio of 0.93. On average, the recombination rate of the combined map was 3.0 cM/Mb. The average recombination rate decreased as the length of the chromosome increased; for the macrochromosomes, a lower recombination rate (about 2 cM/Mb) was observed compared to the microchromosomes (3–14 cM/Mb).

To study the populations separately, linkage maps were calculated for both populations independently (Table 2.2 and 2.3). The linkage map for population 1 consisted of 12,617 markers (95% of the markers used in the combined map) (Additional File 2), and included 306 animals in 42 full- and half-sib families (n =7–13 per family). The number of informative meiosis per mapped marker for population 1 ranged from 20 to 231, with an average of 120. The total length of the sex-average map of population 1 was 3,498.6 cM (Table 2.2). The female sex-specific map was 211.8 cM smaller than the male sex-specific map, with a female-to-male ratio of 0.93. The linkage map of population 2 consisted of 9,803 markers (73% of the markers used in the combined map) (Additional File 2), and included 1,313 animals in 68 full- and half-sib families (n = 6–43 per family). The number of informative meiosis per mapped marker for population 2 ranged from 20 to 1,118, with an average of 551. The total length of the sex-average map of population 2 was 2,812.3 cM (Table 2.3). The female sex-specific map was 198.6 cM smaller than the male sex-specific map, with a female-to-male ratio of 0.93, which was similar to population 1.

Recombination rate

To analyze the recombination frequency along the different chromosomes, the genome was divided into 1,819 nonoverlapping bins with an average size of 560 kb (Additional File 5). For both populations, the sex-average linkage map data were used to calculate the recombination rates of these bins (Figure 2.1). Recombination rates varied from 0 to 60 cM/Mb in population 1 and from 0 to 74 cM/Mb in population 2. Overall, the recombination rates observed between the two populations showed similar trends. Nevertheless, several regions were observed where the two populations diverged with regard to recombination rates (Figure 2.1 and Additional File 5). On GGA 6, 11, 12, and 13, these regions exceeded the stringent Bonferroni threshold when accounting for multiple testing. On these four chromosomes, the regional difference in recombination rate between the two populations was observed at the telomere of the p arm. Similar observations were made in other chromosomes where the two populations diverged with suggestive significance ($p<0.01$).

Table 2.2 The linkage map lengths and recombination rates for the chicken chromosomes of population 1.

Chromosome	Length[1]	Sex-average	Sex-specific		Recombination rate
			Female	Male	
	(Mb)	(cM)	(cM)	(cM)	(cM/Mb)
GGA1	200.9	504	471	541.6	2.5
GGA2	154.8	341.4	321.1	363.5	2.2
GGA3	113.6	288.8	269.5	309.2	2.5
GGA4	94.2	237.6	227.5	247.3	2.5
GGA5	62.2	176.8	175.7	178.5	2.8
GGA6	37.3	110.5	97.9	122.2	3.0
GGA7	38.3	117.1	119.7	118.3	3.1
GGA8	30.6	107.5	103.1	111.3	3.5
GGA9	25.5	97.1	99.0	95.9	3.8
GGA10	22.5	94.5	91.6	97.9	4.2
GGA11	21.9	87.1	86.8	87.7	4.0
GGA12	20.5	89.0	90.3	88.5	4.3
GGA13	18.9	74.1	76.7	71.6	3.9
GGA14	15.8	75.2	74.9	75.4	4.8
GGA15	13.0	59.7	57.0	62.0	4.6
GGA16	0.4	55.4	59.1	53.1	n.d.[2]
GGA17	11.2	54.6	52.4	57.3	4.9
GGA18	10.9	58.1	56.5	60.1	5.3
GGA19	9.9	49.7	52.2	47.9	5.0
GGA20	13.9	58.4	55.8	60.5	4.2
GGA21	6.9	58.9	56.0	61.8	8.5
GGA22	3.9	51.6	55.4	46.5	13.1
GGA23	6.0	48.4	49.1	47.8	8.0
GGA24	6.4	51.2	49.0	53.7	8.0
GGA25	2.0	57.5	56.7	58.5	n.d.[2]
GGA26	5.1	50.6	50.1	50.5	9.9
GGA27	4.7	49.0	47.0	51.3	10.4
GGA28	4.5	52.9	56.8	50.9	11.7
LGE22	0.9	55.6	48.5	62.0	n.d.[2]
LGE64	0.02	23.5	27.4	22.8	n.d.[2]
GGAZ	74.6	262.8	-	262.8	3.5
Total autosomal	956.9	3,235.8	3,133.8	3,355.6	3.4
Total length	1,031.5	3,498.6	3,133.8	3,618.4	3.4

[1] Physical length of the chromosome was based on the position of the last marker in the WASHUC2 build.
[2] n.d.= not determined, as the chromosome showed clear evidence of sequence gaps.

Table 2.3 The linkage map lengths and recombination rates for the chicken chromosomes of population 2.

Chromosome	Length[1]	Sex-average	Sex-specific		Recombination rate
			Female	Male	
	(Mb)	(cM)	(cM)	(cM)	(cM/Mb)
GGA1	200.9	387.1	351.8	428	1.9
GGA2	154.8	267.7	245.9	289.4	1.7
GGA3	113.6	224.8	215.6	236.4	2.0
GGA4	94.2	183.7	171.5	196.5	1.9
GGA5	62.2	148.6	149.1	149.6	2.4
GGA6	37.3	89.8	81.6	97.3	2.4
GGA7	38.3	99.5	93.7	105.2	2.6
GGA8	30.6	94.0	91.9	95.6	3.1
GGA9	25.5	85.2	81.5	88.8	3.3
GGA10	22.5	75.4	73.7	76.3	3.4
GGA11	21.9	58.8	58.4	59.6	2.7
GGA12	20.5	64.3	62.6	66.6	3.1
GGA13	18.9	58.1	58	58.2	3.1
GGA14	15.8	64.2	66.5	61.6	4.1
GGA15	13.0	52.3	51.9	52.4	4.0
GGA16	0.4	0.3	0.5	0.0	n.d.[2]
GGA17	11.2	50.2	51.6	49.3	4.5
GGA18	10.9	49.2	47.8	50.6	4.5
GGA19	9.9	52.7	53.4	52.5	5.3
GGA20	13.9	53.9	54.9	52.9	3.9
GGA21	6.9	56.2	57.2	54.9	8.1
GGA22	3.9	53.6	56	51.9	13.6
GGA23	6.0	53.1	52.2	53.9	8.8
GGA24	6.4	53.7	54.5	52.1	8.4
GGA25	2.0	57.3	54.1	59.4	n.d.[2]
GGA26	5.1	52.6	51.7	53.5	10.3
GGA27	4.7	51.5	52.1	51.3	10.9
GGA28	4.5	53.7	52	55.2	11.9
LGE22	0.9	46.9	54.4	46	n.d.[2]
LGE64	0.02	4.1	4.1	3.8	n.d.[2]
GGAZ	74.6	169.8	-	169.8	2.3
Total autosomal	956.9	2,642.5	2,550.2	2,748.8	2.8
Total length	1,031.5	2,812.3	2,550.2	2,918.6	2.7

[1] Physical length of the chromosome was based on the position of the last marker in the WASHUC2 build.
[2] n.d.= not determined, as the chromosome showed clear evidence of sequence gaps.

The sex-specific linkage maps enabled us to study the effect of sex on recombination. Recombination rates were calculated for nonoverlapping bins based on the recombination rates found in the sex-specific linkage maps of both populations (Figure 2.2 and Additional File 5). Overall, the recombination rates observed between the two sexes of the two populations showed similar trends. However, in the regions on GGA 6, 11, 12 and 13, where the recombination rate of two populations significantly diverged, this difference appeared to be caused by a difference in female recombination rate and not due to male recombination rate (Figure 2.2 and Additional File 5). For the regions where the two populations diverged with suggestive significance ($p<0.01$), the difference in female recombination rate often exceeded the Bonferroni threshold, while there was no statistical evidence for difference in male recombination rate.

2.4 Discussion

The high accuracy of the SNP genotyping, the large number of markers (*n* = 13,340), and the large number of animals (*n* = 1,619) resulted in a high-resolution linkage map of the chicken genome, which significantly exceeds the resolution of previously published linkage maps [7, 10-12]. The current map consists of 13,340 markers, which is an increase of 43.9% compared to the latest consensus map, which comprises 9,268 markers [7]. In total, 2,819 SNP markers overlapped between the two studies. The increased marker density enabled us to efficiently detect genotype errors, thereby increasing the accuracy of the linkage map compared to the latest consensus map.

The use of a large number of animals in the current study resulted in a 6-fold increase (517 vs. 85) in the average number of informative meiosis per mapped marker, thereby increasing the resolution of the current map compared to the latest published linkage map [7]. The higher resolution enabled us, moreover, to order closely linked markers. The linkage map comprises 31 linkage groups, with a total length of 3053.5 cM for the sex-average map of the combined population (Table 2.1). This length is comparable to previous estimates [7].

The construction of separate linkage maps for both populations enabled us to study differences in recombination between the two populations. The sex-average linkage map of population 1 (broiler x broiler cross, 3,498.6 cM) is 24.4% larger than the map of population 2 (purebred broiler line, 2,812.3 cM) (Table 2.2 and 2.3). The difference between the two populations has a biological origin, although differences in informative markers occasionally contributed to the difference between the two maps. An extreme example is GGA16; in population 1, the single

marker located at the end of the chromosome (55.4 cM) was uninformative in population 2, and resulted in a chromosome length of only 0.3 cM in this population. Roughly one third of the difference between the two populations on the autosomal chromosomes is explained by the telomeric regions (defined as 10% of the chromosome length at both telomeres). A clear example is GGAZ, where the difference between the two populations (93 cM) is primarily caused by the telomeric regions. In previous studies, large differences in the length of this chromosome have been reported, varying from 193 to 284 cM [7, 26]. In both populations used in this study, the female specific linkage map was approximately 200 cM smaller than the male specific linkage map. However, for the sex-specific linkage map of population 1, no difference was found between female and male in the latest published linkage map. In addition to having more markers in this study, we also selected more animals and included extra generations of population 1 compared to the last published linkage map. The increased marker density, additional animals, and generations were not expected to have an influence on the (sex-specific) linkage map between the two studies. Nevertheless, the increased number of animals in the current study (and therefore the increased amount of informativity) most likely resulted in a more accurate linkage map, so that the 200 cM difference between female and male recombination could be determined. The 200 cM difference between female and male recombination is, moreover, also seen in population 2, indicating that the female map in chickens is indeed smaller.

Burt and Bell hypothesized that selection leads to high rates of recombination [16]. Although the selection criteria were based on different traits, all three lines used in this study experienced similar selection pressure (personal communication A. Vereijken of Breeding Research and Technology Centre, Hendrix Genetics). We therefore conclude that the difference in recombination between the two populations was not caused by selection pressure *per se*. The linkage map length of the purebred broiler line (population 2) was very similar to that of other chicken populations such as the East Lansing population (partially inbred Red Jungle Fowl x highly inbred White Leghorn cross) and the Uppsala population (Red Jungle Fowl x White Leghorn cross) [7]. Therefore, it appears that the broiler x broiler cross deviates from the other chicken populations by having a high recombination rate. Although not caused by selection, the high recombination rate in this cross could be the result of either a high recombination rate in one or both of the parental lines, or by as-yet unidentified genomic differences between the two lines of this cross.

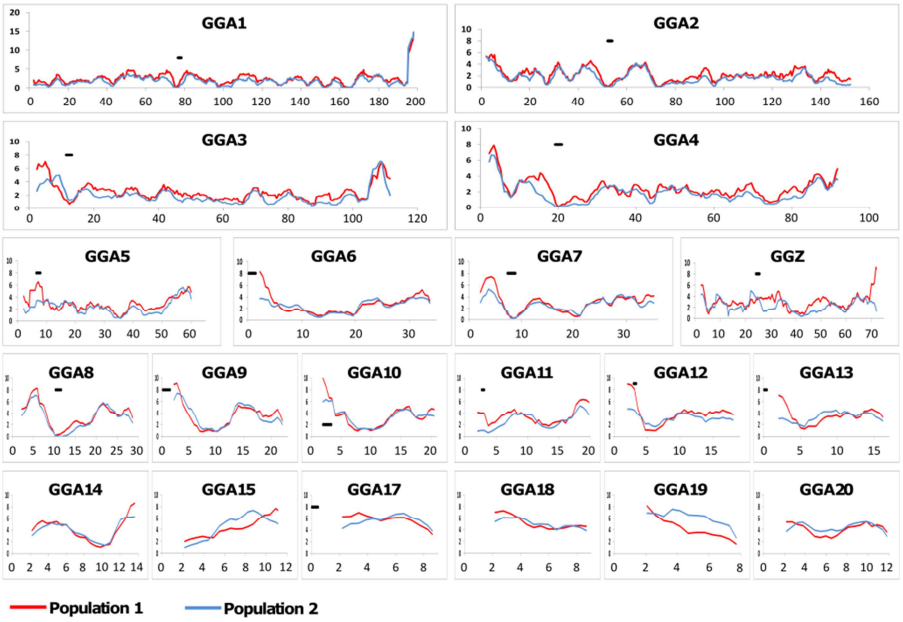


Figure 2.1 Sex-average recombination rate for populations 1 and 2. Recombination rate was calculated for 500 kb nonoverlapping bins, and plotted using a sliding window of eight bins. Population 1 is shown in red and population 2 is shown in blue. On the x-axis, the genomic position is given in million base pairs. On the y-axis, the recombination rate is given in cM/Mb. If known, the position of the centromer is indicated by a solid black line. GGA16, GGA21–GGA28, LGE22, and LGE64 were not included in this figure, because the graphs of these 11 small chromosomes were uninformative. Note that the scale of the y-axis of GGA1 is twice as high as for the other chromosomes.

The high-resolution linkage map enabled us to study recombination hotspots within the two populations and the two sexes (Figures 2.1 and 2.2). Excluding bins with apparent sequence gaps, the recombination rate for the nonoverlapping bins varied from 0 to 20 cM/Mb. This range is in agreement with previous findings in the chicken genome [7]. Overall, recombination rates tended to be similar between the two populations (Figures 2.1 and 2.2). However, when regional differences in recombination hotspots were observed between the two populations, the location of these hotspots were mainly located at the telomere of the p arm (Figure 2.1 and Additional File 5). Moreover, the differences in recombination rate at the telomere appeared to be caused by female-specific recombination hot spots (Figure 2.2 and Additional File 5). Because the broiler x broiler cross (population 1) appears to deviates from other chicken populations, as described above, we conclude that this population had an increased female recombination rate near the telomere of the p arm.

To improve the current genome build, 613 unassigned markers were included on the 18K Illumina iSelect Beadchip. At the time, we assumed that these markers would have a high likelihood of being located on one of the missing microchromosomes, or in sequence gaps that still exist in the current genome build. In total, 59% (*n* = 360) of all unassigned markers were informative in at least one of our two mapping populations. For the markers that had already been mapped to a chromosome, these values were considerably higher: 77% (*n* = 13,250). An explanation for the difference in informativity is that chromosome unassigned is known to be mainly comprised of sequences with lower quality, genome duplications and gene families (e.g. MHC). In particular, genome duplications and gene families are likely to result in the alignment of paralogous sequences, resulting in a higher frequency of false-positive SNPs. These false-positive SNPs contribute to the decreased informativity of the chromosome unassigned markers.

The majority of the informative unassigned SNPs on the beadchip were mapped in sequence gaps of chromosomes or linkage groups that were already covered by the WASHUC2 build. Only 17 SNPs did not appear to be located on any of these chromosomes; however, there was no linkage among these SNPs. The genome coverage for the microchromosomes is, therefore, not extended by the current linkage map. The fact that no new linkage groups were found is in agreement with previous findings that the sequences from the missing chromosomes may be difficult to clone and propagate in *E.coli*; therefore they are missing in the current draft sequence of the chicken genome [7, 13].

In addition to the new markers that were added to improve the current genome build, the high-resolution linkage map presented in this study can be used to correct mistakes in the order of sequences in the current genome assembly. A marker order in the linkage map that is different from the physical map may indicate mistakes in genome assembly. Although the marker order of the linkage map was mainly in agreement with the order of these markers on the physical map, some changes were observed. For the microchromosomes and the two linkage groups, these changes were not unexpected, because several of these chromosomes were known to be poorly assembled. On GGA5 and GGA13, the changed marker order suggests an incorrect genome assembly or a possible inversion in the broiler populations compared to the reference sequence (Red Jungle Fowl). In our data, the inversed marker order on GGA5 led to a 1.6 cM decrease in the length of the population 2, although no reduction in map length was seen in population 1. Similarly, on GGA13, the inversed marker order resulted

in a 1.8 cM decrease in the map of population 2, but had no influence on the map length of population 1.

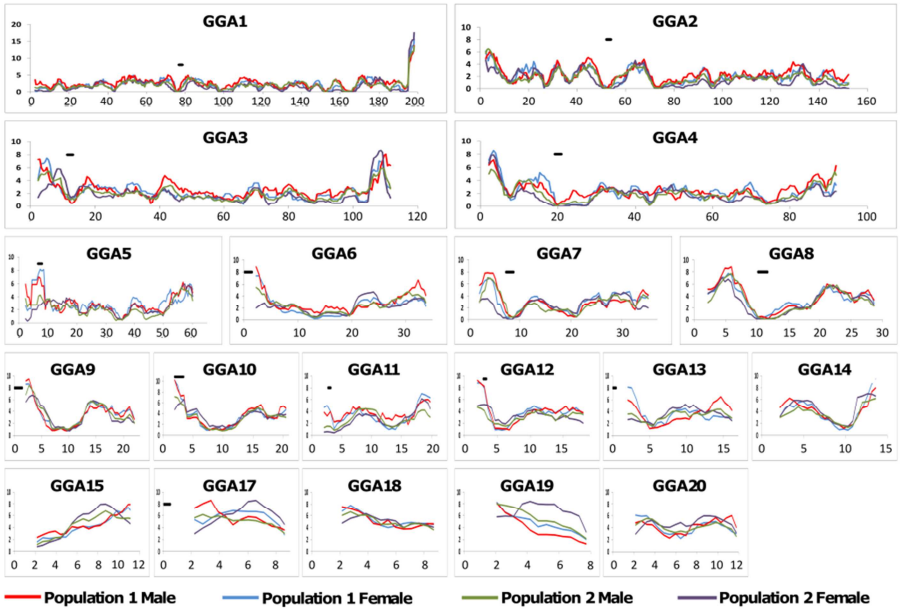


Figure 2.2 Sex-specific recombination rate for populations 1 and 2. Recombination rate was calculated for 500 kb nonoverlapping bins, and plotted using a sliding window of eight bins. The female map of population 1 is shown in blue, and the male map of population 1 is shown in red. The female map of population 2 is shown in purple, and the male map of population 2 is shown in green. On the x-axis, the genomic position is given in million base pairs. On the y-axis, the recombination rate is given in cM/Mb. If known, the position of the centromer is indicated by a solid black line. GGA16, GGA21–GGA28, LGE22, and LGE64 were not included in this figure, because the graphs of these 11 small chromosomes were uninformative. Note that the scale of the y-axis of GGA1 is twice as high as for the other chromosomes.

2.5 Conclusions

In this study, we present a linkage map of the chicken genome at a substantially higher resolution than previously published linkage maps. The increased resolution enabled us to study underlying recombination hotspots. There were regional difference in recombination hotspots between the two mapping populations in several chromosomes near the telomere of the p arm, and sex-specific analysis revealed that these regional differences were caused mainly by female-specific recombination hotspots in the broiler x broiler cross.

Authors' contributions

MG, RC, PA and MGE conceived and designed the experiments. PA, MGE and TV performed the experiments. MGE, PA and TV analysed the data. MGE, MG and RC wrote the paper. All of the authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Addie Vereijken (Breeding Research and Technology Centre, Hendrix Genetics) and Henk Bovenhuis (Wageningen University) for their contribution to this study. This study was part of "The characterisation of genes involved in pulmonary hypertension syndrome in chicken" project funded by the Dutch Technology Foundation (STW). Project number 07106.

Additional files

Additional file 1. Detailed information of all used markers used in the construction of the linkage map.
http://www.biomedcentral.com/content/supplementary/1471-2156-11-11-S1.XLS
Description: This table includes all SNPs used in the construction of the linkage map including their position on the chromosome, newly assigned chromosome if applicable, status and sequence.

Additional file 2. The complete linkage map.
http://www.biomedcentral.com/content/supplementary/1471-2156-11-11-S2.XLS
Description: This file contains the linkage map of the combined and separate population. It includes the sex-average and sex-specific maps.

Additional file 3. Overview of all used markers.
http://www.biomedcentral.com/content/supplementary/1471-2156-11-11-S3.PDF
Description: This figure shows an overview of all markers and includes the number of markers assigned, unassigned, not mapped, mapped, not used, chromosome unassigned, homozygous or rejected by Beadstudio.

Additional file 4. Detailed information about the chromosome unassigned markers.
http://www.biomedcentral.com/content/supplementary/1471-2156-11-11-S4.XLS
Description: This table includes all chromosome unassigned markers, whether they are assigned or mapped to a chromosome, or why they were rejected for the construction of the linkage map.

Additional file 5. The 500kb bins used to study the recombination rates.
http://www.biomedcentral.com/content/supplementary/1471-2156-11-11-S5.XLS
Description: This file includes all 1,819 bins that were used to study recombination rates. It includes the bins for both populations (sex-average, female- and male-specific), and the Z and p-values of the eight bin sliding windows.

References

1. Daw EW, Thompson EA, Wijsman EM: Bias in multipoint linkage analysis arising from map misspecification. *Genetic Epidemiology* 2000, 19(4):366-380.
2. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K: A high-resolution recombination map of the human genome. *Nat Genet* 2002, 31(3):241-247.
3. Kong X, Murphy K, Raj T, He C, White PS, Matise TC: A Combined Linkage-Physical Map of the Human Genome. *Am J Hum Genet* 2004, 75(6):1143-1148.
4. Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S, Chui B, Cohen P, de Toma C, Ehm M, Glanowski S, He C, Heil J, Markianos K, McMullen I, Pericak-Vance MA, Silbergleit A, Stein L, Wagner M, Wilson AF, Winick JD, Winn-Deen ES, Yamashiro CT, Cann HM, Lai E, Holden AL: A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am J Hum Genet* 2003, 73(2):271-284.
5. Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, Mott R, Flint J: A High-Resolution Single Nucleotide Polymorphism Genetic Map of the Mouse Genome. *PLoS Biol* 2006, 4(12):e395.
6. The STAR Consortium : SNP and haplotype mapping for genetic analysis in the rat. *Nat Genet* 2008, 40(5):560-566.
7. Groenen MAM, Wahlberg P, Foglio M, Cheng H, Megens H, Crooijmans RPMA, Besnier F, Lathrop GM, Muir WM, Wong GKS, Gut I, Andersson L: A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Res* 2009, 19:510-519.

8. Bumstead N, Palyga J: A preliminary linkage map of the chicken genome. *Genomics* 1992, 13(3):690-697.
9. Wallis JW, Aerts J, Groenen MAM, Crooijmans RPMA, Layman D, Graves TA, Scheer DE, Kremitzki C, Fedele MJ, Mudd NK, Cardenas M, Higginbotham J, Carter J, McGrane R, Gaige T, Mead K, Walker J, Albracht D, Davito J, Yang SP, Leong S, Chinwalla A, Sekhon M, Wylie K, Dodgson J, Romanov MN, Cheng H, de Jong PJ, Osoegawa K, Nefedov M, Zhang H, McPherson JD, Krzywinski M, Schein J, Hillier L, Mardis ER, Wilson RK, Warren WC: A physical map of the chicken genome. *Nature* 2004, 432(7018):761-764.
10. Jacobsson L, Park H, Wahlberg P, Jiang S, Siegel P, Andersson L: Assignment of fourteen microsatellite markers to the chicken linkage map. *Poult Sci* 2004, 83(11):1825-1831.
11. S. Kerje, Ö. Carlborg, L. Jacobsson, K. Schütz, C. Hartmann, P. Jensen, L. Andersson: The twofold difference in adult size between the red junglefowl and White Leghorn chickens is largely explained by a limited number of QTLs. *Animal Genetics* 2003, 34(4):264-274.
12. Groenen MAM, Cheng H, Bumstead N, Benkel B, Briles W, Burke T, Burt DW, Crittenden L, Dodgson JB, Hillel J, Lamont S, Ponce de Leon A, Soller M, Takahashi H, Vignal A: A Consensus Linkage Map of the Chicken Genome. *Genome Res* 2000, 10:137-147.
13. Hillier L, Miller W, Birney E, Warren W, Hardison R, *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004, 432:695 - 716.
14. Cheng HH, Levin I, Vallejo RL, H K, Dodgson JB, Crittenden L, Hillel J: Development of a genetic map of the chicken with markers of high utility. *Poult Sci* 1995, 74:1855-1874.
15. Schmid M, Nanda I, Burt DW: Second report on chicken genes and chromosomes 2005. *Cytogenet Genome Res* 2005, 109(4):415-479.
16. Burt DW, Bell G: Mammalian chiasma frequencies as a test for two theories of recombination. *Nature* 1987, 326:803-805.
17. International Chicken Genome Sequencing Consortium: A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* 2004, 432(7018):717-722.
18. Infinium genotyping data analysis protocol [http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf]
Accessed September 2007

19. Pakdel A, Van Arendonk JA, Vereijken AL, Bovenhuis H: Direct and maternal genetic effects for ascites-related traits in broilers. *Poult Sci* 2002, 81(9):1273-1279.
20. Closter AM, van As P, Groenen MAM, Vereijken ALJ, van Arendonk JAM, Bovenhuis H: Genetic and phenotypic relationships between blood gas parameters and ascites-related traits in broilers. *Poult Sci* 2009, 88(3):483-490.
21. Rabie TSKM, Crooijmans RPMA, Bovenhuis H, Vereijken ALJ, Veenendaal T, van der Poel JJ, Van Arendonk JAM, Pakdel A, Groenen MAM: Genetic mapping of quantitative trait loci affecting susceptibility in chicken to develop pulmonary hypertension syndrome. *Animal Genetics* 2005, 36(6):468-476.
22. Jennen D, Vereijken A, Bovenhuis H, Crooijmans R, Veenendaal A, van der Poel J, Groenen M: Detection and localization of quantitative trait loci affecting fatness in broilers. *Poult Sci* 2004, 83(3):295-301.
23. van Kaam JBCHM, van Arendonk JAM, Groenen MAM, Bovenhuis H, Vereijken ALJ, Crooijmans RPMA, van der Poel JJ, Veenendaal A: Whole genome scan for quantitative trait loci affecting body weight in chickens using a three generation design. *Livestock Production Science* 1998, 54(2):133-150.
24. van Kaam JBCHM, Groenen MAM, Bovenhuis H, Veenendaal A, Vereijken ALJ, van Arendonk JAM: Whole genome scan in chickens for quantitative trait loci affecting growth and feed efficiency. *Poult Sci* 1999, 78(1):15-23.
25. Documentation for CRI-MAP, version 2.4. [http://linkage.rockefeller.edu/soft/crimap/]
26. Wahlberg P, Strömstedt L, Tordoir X, Foglio M, Heath S, Lechner D, Hellström AR, Tixier-Boichard MH, Lathrop MG, Gut IG, Andersson L: A high-resolution linkage map for the Z chromosome in chicken reveals hot spots for recombination. *Cytogenet Genome Res* 2007, 117:22-29.

3

Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken

M.G. Elferink, A.A.A. Vallée, A.P. Jungerius, R.P.M.A. Crooijmans, M.A.M. Groenen

Animal Breeding and Genomics Centre, Wageningen University and
Research Centre, Marijkeweg 40, PO Box 338, Wageningen, The Netherlands

BMC Genomics (2008) Aug 20; 9: 391

Abstract

One of the loci responsible for feather development in chickens is K. The K allele is partially dominant to the k+ allele and causes a retard in the emergence of flight feathers at hatch. The K locus is sex linked and located on the Z chromosome. Therefore, the locus can be utilized to produce phenotypes that identify the sexes of chicks at hatch. Previous studies on the organization of the K allele concluded the integration of endogenous retrovirus 21 (ev21) into one of two large homologous segments located on the Z chromosome of late feathering chickens. In this study, a detailed molecular analysis of the K locus and a DNA test to distinguish between homozygous and heterozygous late feathering males are presented.

The K locus was investigated with quantitative PCR by examining copy number variations in a total of fourteen markers surrounding the ev21 integration site. The results showed a duplication at the K allele and sequence analysis of the breakpoint junction indicated a tandem duplication of 176,324 basepairs. The tandem duplication of this region results in the partial duplication of two genes; the prolactin receptor and the gene encoding sperm flagellar protein 2. Sequence analysis revealed that the duplication is similar in Broiler and White Leghorn. In addition, twelve late feathering animals, including Broiler, White Leghorn, and Brown Layer lines, contained a 78 bp breakpoint junction fragment, indicating that the duplication is similar in all breeds. The breakpoint junction was used to develop a TaqMan-based quantitative PCR test to allow distinction between homozygous and heterozygous late feathering males. In total, 85.3% of the animals tested were correctly assigned, 14.7% were unassigned and no animals were incorrectly assigned.

The detailed molecular analysis presented in this study revealed the presence of a tandem duplication in the K allele. The duplication resulted in the partial duplication of two genes; the prolactin receptor and the gene encoding sperm flagellar protein 2. Furthermore, a DNA test was developed to distinguish between homozygous and heterozygous late feathering males.

3.1 Introduction

One of the loci responsible for feather development in chickens was described by Serebrovsky in 1922 [1] and is designated by the symbol K, standing for 'kürzer flügel' (short wing) [2]. The K allele is associated with the late feathering phenotype (LF) that causes a retard in the emergence of primary and secondary flight feathers. The k+ allele is associated with the early feathering phenotype (EF), resulting in the earliest emergence of feathers. The K allele appears to be incompletely dominant to k+, resulting in phenotypes with different intensities due to a dosage effect of the locus [3]. For more detailed information about the feathering loci, see the extensive review by Chambers *et al.* [4].

In birds, sex is determined by two chromosomes, Z and W. Males are homozygous ZZ and females are hemizygous ZW. The K locus is located on the Z chromosome and can be utilized to produce phenotypes that distinguish between the sexes of chicks at hatching, but also at the embryonic stage [5, 6]. This method of sexing based on differences in the rate of feather growth provides a convenient and inexpensive approach.

Although the LF phenotype facilitates the sexing of chicks, the K allele is also associated with a reduction in egg production, an increase in infection by lymphoid leucosis virus [7], and an increase in the mortality rate [8]. These negative side effects may be caused by the presence of the endogenous retrovirus 21 (ev21) [8]. Concordance between expression of ev21 and the LF phenotype indicated a linkage of less than 0.3 cM between K and the ev21 locus [9, 10]. The ev21 locus consists of an integration site that can be occupied (ev21+) or unoccupied (ev21-). EF animals were found to have only one unoccupied site per Z chromosome; whereas, LF animals have at least one Z chromosome with an unoccupied and an occupied site [11]. A study on the organization of the K allele concluded the integration of ev21 into one of two large homologous segments located on the Z chromosome of LF chickens [12]. EF revertants carrying an occupied site have been observed; therefore, it was concluded that ev21 itself could not be the sole cause of the LF phenotype [13].

Several tests have been developed to identify the EF and LF alleles [12, 14, 15]. These tests focused on the presence of the occupied and unoccupied site in the genome. Unfortunately, even if these methods are fully informative when applied to females, they do not allow for differentiation between homozygous and heterozygous males. Furthermore, the existence of ev21-positive EF animals will give false-positive results with these tests.

In this study we present a detailed molecular analysis of the K locus and develop a DNA test to distinguish between homozygous and heterozygous late feathering males.

3.2 Results

Molecular analysis of the K locus

A quantitative PCR (qPCR) approach, as described by Weksberg *et al.* [16], was used to investigate the K locus. Copy number variation was determined at fourteen markers (STS_1-STS_14) designed to surround the ev21 integration site (Table 3.1). In two chickens, the most likely location of the duplicated block was mapped between markers STS_6 and STS_13 (Table 3.2). Marker STS_5 and marker STS_6 gave ambiguous results (Table 3.2).

To determine the size and orientation of the duplicated block, forward and reverse primers were designed for both ends (between marker STS_6 and STS_7, and between markers STS_13 and STS_14). A 1238 bp product was obtained spanning the breakpoint junction (marker STS_junction) in two late feathering males. With this marker, no PCR product was obtained from the DNA of the two EF birds. Sequence analysis of the PCR product obtained from the two LF males provided the exact breaking point. Based on the WASHUC2 assembly, the total length of the tandem duplication is 176,324 bp (GGAZ 9,966,364-10,142,688 bp). The tandem duplication of this region results in the partial duplication of two genes: the prolactin receptor (*PRLR*) and the gene encoding sperm flagellar protein 2 (*SPEF2*, also known as *KPL2*). The duplicated block included exons 1 to 11 and 558 bp of exon 12 of *PRLR*, and exons 1 to 5 of *SPEF2* (Figure 3.1). No differences in the nucleotide sequences of the breakpoint junction fragments were observed between the Broiler and White Leghorn animals.

To validate the duplication, a PCR reaction was performed with a new marker spanning the breakpoint junction (STS_break). The experiment was performed on twelve EF and twelve LF animals from eight different lines. No band was observed for the EF animals; whereas, all LF animals showed the 78 bp band corresponding to the breakpoint junction.

To obtain information about possible aberrations at the ends of the duplication, both regions were sequenced (markers STS_5block and STS_3block). No sequence differences were found between the LF and wildtype (EF) animals.

Table 3.1 STS markers used in the molecular analysis of the K locus.

Marker Name	Location[1] (bp)	Position	Sequence	Length (bp)
STS_0	80092619 [2]	Forward	CACACAGAAGACGGTGGATG	170
	80092788 [2]	Reverse	TGGCTCCTACCTCCTGACAC	
STS_1	9764119	Forward	GAAGGAGAGCCTGTTTGCTG	207
	9764325	Reverse	CTTGTGGTGGTGAAGTGGTG	
STS_2	9862778	Forward	AAGTGGGACAACGGAAAGAC	345
	9863122	Reverse	AGGTCAAAGAAGGCACAAGG	
STS_3	9913200	Forward	AGCCAGAAACAAAAGCCAAA	148
	9913347	Reverse	TCAGCCTCGACACAGAAAAA	
STS_4	9933229	Forward	AGTGTCAGTGTGCCTCTTGG	170
	9933398	Reverse	CACGGCATTTATGAGATTGG	
STS_5	9950543	Forward	AATCAGAGTTGCAGGGGTTG	135
	9950677	Reverse	TTGACTGGGGCTCAATAAGG	
STS_6	9960545	Forward	TCTCCCTCCCTGTCTTCTCA	215
	9960759	Reverse	TGGCCTTGAAAATCCTCTTG	
STS_7	9973781	Forward	TAGCAGACAAGGGCATTCAG	198
	9973584	Reverse	GCATTTGTAGGGCTGGATTTG	
STS_8	9996871	Forward	ACCAAAGCGTCCAAAATGTC	198
	9997068	Reverse	TACCAGGGGAGAGCATGAAG	
STS_9	10038160	Forward	AAATAGGCACGAGGGAAGC	176
	10037985	Reverse	AACCATCAAGACTGGCTCAAC	
STS_10	10078039	Forward	GCCCTCTAAGTGCCTGACTG	182
	10078220	Reverse	TTTCATGCGTAGGAGCTGTG	
STS_11	10106858	Forward	CACTTCCAGGGTTGGTGACT	343
	10107200	Reverse	GAGGGCATCCATCACATCTC	
STS_12	10135701	Forward	TGGAGCTGAGGAAAGAATCC	105
	10135805	Reverse	TGCTTGCAGGTTTGAGTGTC	
STS_13	10168014	Forward	TCCACTTGTCATGCACTTCC	179
	10168192	Reverse	AAGTTCCCCAAAAATACTGCTG	
STS_14	10181226	Forward	TGTGAGCAATTCCATTCTGG	216
	10181441	Reverse	TTGGTTCAGTTTGGTCATCG	
STS_Junction	10141819	Forward	CTGAGAGTGTTGTCCCAGCA	1432 [3]
	9966922	Reverse	TGTTGAGTGCTCTTGGTTGC	
STS_Control	9899810	Forward	ACGCTGGCTTTCCCAACAG	70
	9899879	Reverse	AGACTGCCACATACCAGAAGCA	
STS_Break	10142644	Forward	ACAAGTGTCAGACTAGGGCTAGCA	783
	9966396	Reverse	TGAAACCATCCCTGGAGAGATG	
STS_5block	9965590	Forward	ACCATTTCCACATTCCCTTCT	1333
	9966922	Reverse	TGTTGAGTGCTCTTGGTTGC	
STS_3block	10141819	Forward	CTGAGAGTGTTGTCCCAGCA	1289
	10143107	Reverse	CGGGCCATTATTTCATTTTG	

[1]) Genomic location on the Z-chromosome in basepairs (WASHUC2 assembly).
[2]) Marker STS_0 is located on chromosome 1.
[3]) In late feathering animals only.

DNA test to distinguish between homozygous and heterozygous late feathering males

The breakpoint junction was used to develop a TaqMan-based DNA test that can distinguish between homozygous and heterozygous LF males (further referred to as the TaqMan K test). Two TaqMan markers were used: one outside the duplicated block (marker STS_control) was used as a control and one spanning the breakpoint junction (marker STS_break) was used for investigating the duplication (Table 3.1). Two minor groove binding (MGB)-probes were designed for these markers, the MGB-control probe (TCTGTCCAAACATTTATTTG) was labeled with the fluorescent dye VIC and used for the control marker STS_control, and the MGB-Break probe (CCCTTAAATGCCTTGCTT) was labeled with the fluorescent dye FAM and used for the breakpoint junction marker STS_break**.**

To validate the TaqMan K test, 25 animals were tested in duplicate. Eight randomly selected reference animals (four K/K and four K/k+) were used to determine the range of K/K and K/k+ animals in each experiment (Table 3.3). Seventeen animals with known genotypes were used to validate the ranges (Table 3.4). In the first experiment, an animal was considered K/K if the ΔCt was between 0.68 and 1.43 or K/k+ if the ΔCt was between 1.75 and 2.50. For the second experiment, the range of ΔCt for K/K was between 0.63 and 1.24 and between 1.50 and 2.10 for K/k+. Based on these calculations, 94.1% of the animals in the first experiment were within the ranges of their known genotype (correctly assigned), and 5.9% were outside either range (unassigned). No animals were false positive (incorrectly assigned). In the second experiment, 76.5% of the animals were correctly assigned, 23.5% were unassigned and no animals were incorrectly assigned. In total, 29 of the 34 validation animals (85.3%) were correctly assigned, 5 animals (14.7%) were unassigned and no animals were incorrectly assigned.

Table 3.2 The ΔKCt values for the STS markers in two chickens.

Breed[1]	Sex	STS_1	STS_2	STS_3	STS_4	STS_5	STS_6	STS_7
BR	Male	0.05	0.20	0.21	0.11	*0.40*	0.20	**1.17**
WL	Male	-0.03	-0.04	0.33	-0.04	0.01	*0.38*	**1.24**

Breed[1]	Sex	STS_8	STS_9	STS_10	STS_11	STS_12	STS_13	STS_14
BR	Male	**1.49**	**1.52**	**1.71**	**1.19**	**1.36**	0.29	0.14
WL	Male	**1.11**	**1.21**	**1.62**	**1.13**	**1.23**	0.13	0.15

[1] BR: Broiler, WL: White Leghorn. Normal font indicates a ΔKCt ≤0.35.
An italic font indicates a ΔKCt >0.35 and <0.65. A bold font indicates a ΔKCt ≥0.65.

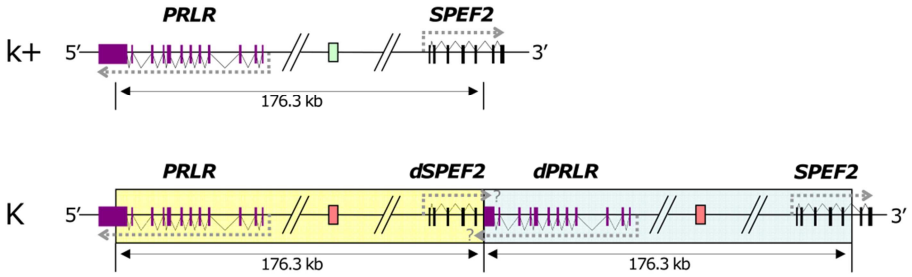


Figure 3.1 The organization of the k+ and K alleles. The k+ allele contains two genes; *PRLR* (purple exons) and *SPEF2* (black exons). The K allele contains the original genes and the two partial duplicate genes, *dPRLR* and *dSPEF2*. The green box indicates the unoccupied ev21 integration site. One of the red boxes indicates an unoccupied and the other an occupied ev21 integration site. The large yellow and blue boxes indicate the 176.3kb duplicated block. The grey arrows indicate transcriptional start and stop. The question mark indicates a transcript of unknown length.

3.3 Discussion

The detailed molecular analysis presented in this study confirmed the presence of the duplication first described by Iraqi and Smith [12]. The total size of the tandem duplication is 176,324 bp, which is in agreement with the estimated 180kb [12]. Sequence analysis found that the duplication is similar in both Broiler and White Leghorn lines, and all 12 LF animals showed the 78 bp breakpoint junction fragment (marker STS_break in the current study) indicating that the duplication is similar in all animals. This suggests that the duplication was of the same origin for all three breeds, and that the duplication most likely occurred in a common ancestor. On the other hand, since the K allele is extensively used by breeders, it is also likely that this particular allele was introduced into all three breeds.

In theory, the values of unaffected and duplicated markers should be equal to 0 or 1, respectively, in the qPCR experiments. However, ΔKCt varied from -0.04 to 1.71, and markers STS_5 and STS_6 had ambiguous results (Table 3.2). This variation is likely to be due to biological variations and the fact that the experiment was only performed once with two animals.

The observed duplication could be the result of an unequal recombination event in the Z chromosome. However, no apparent sequence homologies are found in the two areas involved in the duplication. Therefore, the unequal recombination event is not supported by our data, although a nonhomologous recombination event cannot be excluded. Alternatively, integration of ev21 resulted in the duplication at the K locus. This raises the possibility of additional duplications at other locations in the chicken genome, which contains approximately 12,000 copies of long terminal

repeats (1.3%) belonging to the vertebrate-specific class of retroviruses [17]. However, the actual ends of the duplicated block are located approximately 70kb upstream and 103kb downstream of the ev21 integration site, making this possibility less likely.

A PCR amplicon spanning the breakpoint junction is sufficient for distinguishing LF birds from EF birds. In males however, the challenge was to be able to differentiate between LF homozygous (K/K) and LF heterozygous (K/k+) animals. In this study, we found that the duplicated block is specific for the K allele and it was used to develop a DNA test based on the breakpoint junction. Since the PCR reactions in the TaqMan K test are performed in a multiplex, the concentration of DNA, theoretically, has no influence on the ΔCt. This contributes to the robustness of the test since variations in the concentration of DNA between and within test and control animals does not have an influence on the results. The ΔCt value gives an indication of the haplotype of an animal. In theory, when ΔCt is equal to 1, the animal is heterozygous, and when ΔCt is equal to 0, the animal is homozygous (Figure 3.2). In the TaqMan K test experiments, the homozygous reference animals had an average ΔCt of 1.06 and 0.94, and the heterozygous reference animals had an average ΔCt of 2.13 and 1.80 (Table 3.3). This difference from the theoretical value was most likely caused by the different efficiencies of the markers.

Table 3.3 The TaqMan-based DNA test for the K allele on reference animals.

Animal ID	Genotype	Experiment 1	Experiment 2
		Δ Ct	Δ Ct
6333	K/K	0.92	0.79
4148	K/K	1.14	0.77
4384	K/K	1.16	1.13
6323	K/K	1.00	1.05
949	K/k+	2.15	1.76
6182	K/k+	2.09	1.62
2636	K/k+	1.90	1.66
947	K/k+	2.38	2.14
Average	K/K	1.06	0.94
	K/k+	1.80	2.13

The aim was to develop a highly reliable test that is convenient for intensive use. The reliability of the test was defined by the percentage of correctly and incorrectly assigned animals. The TaqMan K test was validated using eight reference and seventeen validation animals in duplicate. Of the validation animals tested, 85.3% were identified correctly, 14.7% were unassigned, and no animals were incorrectly assigned (Table 3.4). Based on the literature, no previous test has been capable of identifying LF homozygous and LF heterozygous males with this level of reliability.

Although the LF phenotype facilitates the sexing of chicks at hatching, expression of ev21 is associated with the negative side effects of the K allele [7,8]. The establishment of a line where late-feathering is not associated with decreased egg production and tolerance to exogenous avian leucosis virus infection would be of prime commercial interest. Obviously, the search for the K allele lacking the occupied site is an effective approach. This search for revertants and the establishment of a line can be done by combining the TaqMan K test and the ev21 test proposed by Tixier-Boichard [15].

The observed duplication resulted in the partial duplication of two genes: *PRLR* and *SPEF2* (Figure 3.1). The genes are oriented in opposite directions; therefore, the duplication event does not result in a fusion gene. However, alternative transcripts of the partially duplicated genes may be found. Interestingly, the transcript of both partially duplicated genes could contain the antisense sequence of the other gene, which could lead to RNA interference and influence the translation of both the duplicated and original genes.

The membrane-bound PRLR is closely related to the growth hormone receptor and is a member of the cytokine receptor family [18]. The pituitary hormone, prolactin (PRL), is a ligand of PRLR. More than 300 separate biological activities have been attributed to PRL: reproduction, endocrine signaling and metabolism, control of water and electrolyte balance, growth and development, neurotransmission and behavior, and immunoregulation and protection [19]. More detailed functions of PRL include involvement in the control of seasonal pelage cycles [20, 21, 22], egg production [23], and the induction of molting [24]. Furthermore, PRL is involved in the immune system [25], autoimmune diseases, and the growth of different forms of cancer [18]. In *PRLR* (-/-) knockout studies on mice, the normal progression of hair replacement and follicle development have been observed [26]. These knockout mice showed a change in the timing of hair replacement and molting, and both phenotypes are advanced compared to the wild type. It was concluded that knocking out *PRLR* shortens the telogen phase of the hair cycle and advances the anagen phase of hair follicles [26, 27]. Therefore, it can be suggested that PRLR plays an inhibitory role in follicle activation.

The relatively unknown protein, SPEF2, is believed to play an important role in the differentiation of axoneme-containing cells [28]. Truncation of the SPEF2 protein results in immotile short-tail sperm in pigs [29]. Due to the presence of an ATP/GTP binding site and a proline rich domain, it was suggested that SPEF2 might be involved in signal transmission [28].

Table 3.4 The TaqMan-based DNA test for the K allele validated on late feathering K/K and K/k+ animals.

Animal ID	Known	Experiment 1		Experiment 2	
	Genotype	ΔCt	Genotype	ΔCt	Genotype
2864	K/k+	0.76	K/k+	0.64	K/k+
B2L4	K/k+	0.68	K/k+	0.49	unassigned
942	K/k+	0.90	K/k+	1.01	K/k+
2855	K/k+	0.98	K/k+	0.87	K/k+
4117	K/k+	1.10	K/k+	0.40	unassigned
4118	K/k+	0.98	K/k+	0.83	K/k+
4332	K/k+	1.31	K/k+	1.14	K/k+
6388	K/k+	1.06	K/k+	0.77	K/k+
6324	K/k+	1.12	K/k+	0.91	K/k+
6130	K/K	2.44	K/K	1.84	K/K
6297	K/K	2.09	K/K	1.40	unassigned
952	K/K	2.09	K/K	1.74	K/K
1030	K/K	1.83	K/K	1.9	K/K
2849	K/K	2.26	K/K	1.64	K/K
6187	K/K	2.10	K/K	1.85	K/K
6242	K/K	1.73	unassigned	1.50	K/K
6172	K/K	1.93	K/K	1.47	unassigned

	Experiment 1		Experiment 2		Total	
	Animals (n=17)	%	Animals (n=17)	%	Animals (n=34)	%
Correct	16	94.1	13	76.5	29	85.3
Incorrect	0	0	0	0	0	0
Unassigned	1	5.9	4	23.5	5	14.7

The seventeen animals were validated based on the ranges found for K/K and K/k+. For experiment 1, the Δ Ct range for K/K was 0.68-1.43 and for K/k+ 1.75-2.50. For experiment 2, the Δ Ct range for K/K was 0.63-1.24 and for K/k+ 1.50-2.10.

The actual cause of delayed feathering is still unknown. It can be speculated that *PRLR*, due to its inhibitory role in follicle activation, is the major candidate gene involved in this delay. *SPEF2* may be involved in the transmission of signals in the feather growth pathway. Further research is needed to confirm the involvement of these genes, which could focus on 1) the truncated proteins formed by *PRLR* or *SPEF2* as a result of the partial duplication, 2) the transcripts of the partially duplicated genes and their influence on the expression and translation of the two original genes, and 3) the expression of (partially duplicated) *PRLR* and *SPEF2* that may have changed due to the rearrangement, duplication, or deletion of regulatory elements.

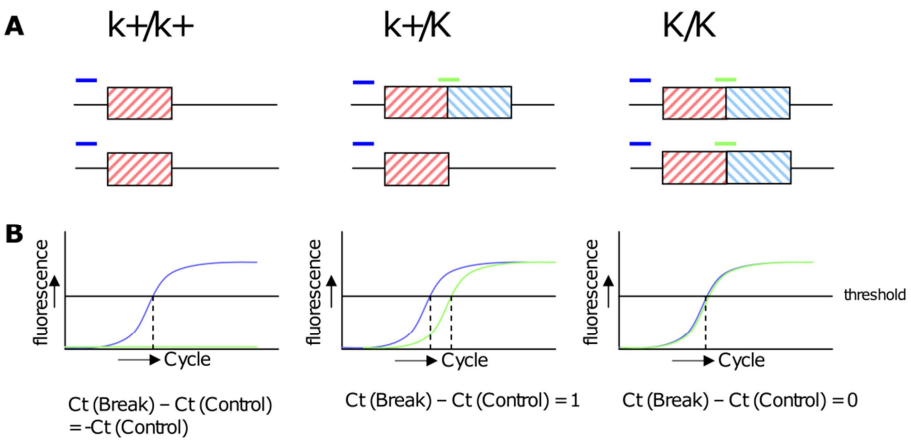


Figure 3.2 Difference in the Ct values of homozygous early feathering (EF), heterozygous late feathering (LF), and homozygous LF animals. A) Comparison of K locus components on the Z chromosomes of different genotypes. The red striped box and blue striped box indicate the duplicated blocks of genetic sequence. The dark blue line is marker STS_control and the green line is marker STS_break. B) The theoretical curves of the qPCR. In k+/k+ animals the difference between Ct (Break) and Ct (Control) will be –Ct (Control). For k+/K animals the theoretical difference will be 1 Cycle. For K/K animals the difference will be 0.

Although it has been extensively described that ev21 causes the negative side effects of the K allele, the findings of this study might also indicate involvement of *PRLR*. As described above, prolactin and its receptor are involved in the growth of different forms of cancer [18], egg production [23], and in the immune system [25]. Because the negative side effects of the K allele include an increase in infection by lymphoid leucosis virus, an increased mortality, and a reduction in egg production, it can be speculated that the partial duplication, altered expression, or altered translation of *PRLR* might also be involved in the negative side effects. If the partial

duplication of *PRLR* is responsible for the delay in feather growth, and contributes to the negative side effects, it will not be possible to separate the advantageous and disadvantageous effects of the K allele.

3.4 Conclusions

The detailed molecular analysis presented in this study indicates the presence of a 176,324 bp tandem duplication in the K allele. An identical duplicated block is found in Broiler, White Leghorn, and Brown Layer lines. The duplication results in the partial duplication of two genes: *PRLR* and *SPEF2*. Due to its inhibitory role in follicle activation, *PRLR* is the most likely candidate gene involved in the delay of feather growth. However, *SPEF2* may be involved in the transmission of signals in the feather growth pathway.

In addition to the characterization of the K locus, a DNA test was developed to distinguish between homozygous and heterozygous late feathering males. The percentage of animals correctly assigned was 85.3%, while 14.7% were unassigned. No animals were incorrectly assigned. To date, this is the most reliable and robust DNA test developed to differentiate between LF homozygous and LF heterozygous males, and would be indispensable in decreasing errors generated by crossing animals with incorrect genotypes.

3.5 Methods

DNA collection

Chicken genomic DNA was extracted from the blood of EF and LF animals provided by Hendrix Genetics (the Netherlands) using the Puregene DNA purification blood kit (Gentra System, USA). DNA concentration and quality were measured using the Nanodrop ND-1000 spectrophotometer. In total, 14 homozygous EF males (k+/k+), 23 homozygous LF males (K/K), three LF females (K/W), and 12 heterozygous LF males (K/k+) from three different lines (Broiler, White Leghorn, and Brown Layer) were used. The genotypes were determined by examining the feathering phenotypes of their offspring.

Primers and probes

The TaqMan primers and probes were designed using Primer Express 3.0 (Applied Biosystems) and all other primers were designed using Primer3 [30]. All primers were designed using sequence information from assembly WASHUC2 (may 2006), available on the Ensembl website [31].

Molecular analysis of the K locus

For the 15 STS markers (STS_0 to STS_14), the criteria for primer design were as follows: amplicons of 100 to 250bp, primer melting temperature ranging from 58°C to 62°C, primer length ranging from 19 to 22 bp, and primer G/C content ranging from 40% to 60%. Slope values were calculated using software from Applied Biosystems (SDS1.2) and an input of 50, 5, 0.5, and 0.05 ng (10^2 - 10^{-2}) DNA was used in duplicate. The slope values of all markers were within the range of -3.32 ± 0.25 [16] and the R^2 of all markers was above 0.994. Marker STS_0, designed in the glyceraldehyde-3-phosphate dehydrogenase gene, was used to normalize the data. The qPCR experiment was performed with the Real-time PCR 7500 from Applied Biosystems. Each 25 µl qPCR reaction was comprised of 12.5 µl IQ SYBR GREEN mastermix (Biorad), 300 nM of each primer, and 20 ng of genomic DNA. Genomic DNA from two EF (one Broiler and one White Leghorn) and two LF animals (one Broiler and one White Leghorn) were tested once for all markers. The PCR program was 50°C for 2 min, a 10 min denaturation at 95°C, then 40 cycles of 95°C for 15 sec and combined annealing and extension at 60°C for 60 sec. At the end, a dissociation step was included to confirm the specificity of the product. Results were expressed in the number of cycles (Ct value) at a threshold of 100,000 ΔRn. The method described by Sijben *et al.* [32] was used to normalize the Ct values (KCt). All data was normalized against the Ct values of marker STS_0. Slope values were included in the calculations.

For all markers, the average KCt was calculated for both EF animals and substracted from the KCt of each LF animal (ΔKCt). When the ΔKCt of a marker was less than 0.35, no duplication was observed; when ΔKCt was between 0.35 and 0.65, the result was ambiguous and no conclusion could be given; and when ΔKCt was more than 0.65, it indicated a gain of one copy and, therefore, a duplicated marker [16].

In order to obtain the exact breakpoint, and to identify specific SNPs in this region, the PCR reaction was performed on one EF male and one LF male from two breeds (Broiler and White Leghorn). The PTC-100 Thermal Controller (MJ Research, Inc.) was used. The PCR reaction (10 µl total volume) was comprised of 5 µl ABgene PCR mastermix, 400 nM of each primer, and 20 ng of genomic DNA. The PCR program was 95 °C for 5 min, followed by 36 cycles of 95 °C for 30 sec, 60 °C for 45 sec, and 72 °C for 1 min 30 sec, with a final extension at 72 °C for 10 min. Amplified products were separated at 115 V for 45 min on a 1.5% agarose gel. The products of marker STS_Junction, STS_5Block, and STS_3Block were amplified and sequenced using the Applied Biosystems 3730 DNA analyzer. The standard protocol of the Big Dye Terminator Cycle Sequencing Kit v3.1 (ABI) was used. Sequence data was analyzed using Pregap4 and Gap4 of the Staden Software Package [33]. The Pregap4

modules were used to prepare the sequence data for assembly (quality analysis). Gap4 was used for the final sequence assembly of the Pregap4 output files (normal shotgun assembly).
In addition, PCR reactions were performed on the breakpoint junction in twelve EF and twelve LF animals using the breakpoint junction marker STS_break (Table 3.1). Eight different lines were used: four EF and four LF lines consisting of four Broiler, two White Leghorn, and two Brown Layer lines. From each line, three animals were used in the experiment. The three LF White Leghorn animals were female. The PCR method was similar to that described above.

The TaqMan K test

Standard curves were generated using the SDS1.2 software from Applied Biosystems with a DNA concentration of 5, 0.5, and 0.05 ng in triplicate. Marker STS_control had a R^2 value of 0.995 and a slope of -3.36. Marker STS_break had a R^2 of 0.977 and a slope of -4.31. For marker STS_break, no marker could be developed with a higher R^2 or a higher slope. Each 25 µl qPCR reaction was comprised of 12.5 µl ABgene PCR master mix, 300 nM of each primer, 100 nM of each probe, and 5 ng genomic DNA. The breakpoint junction and control primers and probes were used in multiplex within one reaction. The experiments were performed using the same PCR program used in the qPCR experiments, but without a dissociation step. Based on the results, the threshold was kept at 9200 ΔRn for all calculations. The difference in the number of cycles between the breakpoint junction and control marker was calculated (ΔCt = Ct FAM – Ct VIC). The difference between the average ΔCt of eight reference animals (four K/K and four K/k+) was used to calculate the DΔCt (DΔCt = ΔCt K/K - ΔCt K/k). This DΔCt was then used to calculate a range of ΔCt values to distinguish between K/K and K/k+ (Figure 3.3). An animal was assigned as homozygous (K/K) if the ΔCt was in the range of -35% to +35% DΔCt of the average from the homozygous reference animals. An animal was assigned as heterozygous (K/k+) if the ΔCt was in the range of -35% or +35% DΔCt of the average from the heterozygous reference animals. The ΔCt values outside these ranges were considered to be unassigned and when a tested animal was placed into the wrong genotype it was considered to be incorrectly assigned (false positive).

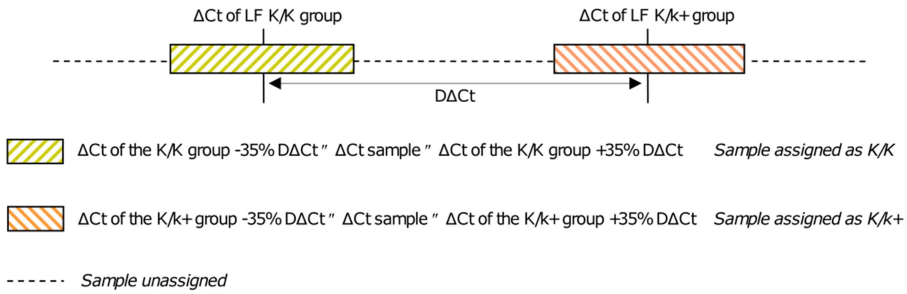


Figure 3.3 Range of ΔCt used to identify the genotype of the tested animals.

Authors' contributions

MGE and AV drafted the manuscript and designed, conducted, and analyzed the experiments. AJ, RC, and MG participated in the design of the experiments and helped substantially with manuscript preparation and editing. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Tineke Veenendaal and Kaveh Hemmatian for their excellent help and guidance in conducting the experiments. Gerard Albers and Addie Vereijken from the Breeding Research and Technology Centre (Hendrix Genetics) are thanked for providing blood samples and for their contribution to this study. This study was funded by the Euribrid Breeding Research Centre Boxmeer (Hendrix Genetics), the Netherlands.

References

1. Serebrovsky AS: Crossing-over involving three sex-linked genes in chickens. Amer Nat 1922, 56:571-572.
2. Hertwig P, Rittershaus T: Die Erbaktoren der Haushuhner. Z ind Abst Vereb 1929, 51:354-72.
3. Siegel PB, Mueller CD, Craig JV: Some phenotypic differences among homozygous, heterozygous, and hemizygous late feathering chicks. Poult Sci 1957, 36:232-239.
4. Chambers JR, Smith EJ, Dunnington EA, Siegel PB: Sex-linked feathering (K, k+) in chickens: a review. Poult Sci 1993, 5:97-116.
5. Radi MH, Warren DC: Studies on the physiology and inheritance of feathering in the growing chick. J Agric Res 1938, 56:679-705.
6. Warren DC: Developing early-feathering strains in heavy breeds of poultry. Agricultural experiment station 1944, 224.

7. Harris DL, Garwood VA, Lowe PC, Hester PY, Crittenden LB, Fadly AM: Influence of sex-linked feathering phenotype of parents and progeny upon lymphoid leucosis virus infection status and egg production. Poult Sci 1984, 63(3):401-413.
8. Smith EJ, Fadly AM: Influence of congenital transmission of endogenous virus 21 on the immune response to avian leucosis virus infection and the incidence of tumors in chickens. Poult Sci 1988, 67:1674-1679.
9. Bacon LD, Smith E, Crittenden LB, Havenstein GB: Association of the slow feathering (K) and an endogenous viral (ev21) gene on the Z chromosome of chickens. Poult Sci 1988, 67(2):191-197.
10. Smith EJ, Fadly AM: Male-mediated venereal transmission of endogenous avian leucosis virus. Poult Sci 1994, 73(4):488-494.
11. Boulliou A, Le Pennec JP, Hubert G, Donal R, Smiley M: The endogenous Retroviral ev21 locus in commercial chicken lines and its relationship with the slow-feathering phenotype K. Poult Sci 1992, 71:38-46.
12. Iraqi F, Smith EJ: Organization of the sex-linked late-feathering haplotype in chicken. Anim Genet 1995, 26:141-146.
13. Levin I, Smith EJ: Molecular analysis of endogenous virus ev21-slow feathering complex of chickens. 1. Cloning of proviral-cell junction fragment and unoccupied site. Poult Sci 1990, 69(11):2017-2026.
14. Smith EJ, Levin I: Application of a locus-specific DNA hybridization probe in the analysis of the slow-feathering endogenous virus complex of chickens. Poult Sci 1991, 70(9):1957-1964.
15. Tixier-Boichard MH, Benkel BF, Chambers JR, Gavora JS: Screening chickens for endogenous virus ev21 viral element by the polymerase chain reaction. Poult Sci 1994, 73(10):1612-1616.
16. Weksberg R, Hughes S, Moldovan L, Bassett AS, Chow EW, Squire JA: A method for accurate detection of genomic microdeletions using real-time quantitative PCR. BMC Genomics 2005, 6:180.
17. Hillier L, Miller W, Birney E, Warren W, Hardison R, et al.: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 2004, 432:695-716.
18. Bole-Feysot C, Goffin V, Edery M, Binart N, Kelly PA: Prolactin (PRL) and its receptor: actions, signal transduction pathways and phenotypes observed in PRL receptor knockout mice. Endocr Rev 1998, 19(3):225-68.
19. Goffin V, Bernichtein S, Touraine P, Kelly PA: Development and potential clinical uses of human prolactin receptor antagonists. Endocr Rev 2005, 26(3):400-422.

20. Martinet L, Allain D: Role of the pineal gland in the photoperiodic control of reproductive and non-reproductive functions in mink (Mustela vison). Ciba Found Symp 1985, 117:170-187.
21. Pearson AJ, Parry AL, Ashby MG, Choy VJ, Wildermoth JE, Craven AJ: Inhibitory effect of increased photoperiod on wool follicle growth. J Endocrinol 1996, 148(1):157-166.
22. Nixon AJ, Ford CA, Wildermoth JE, Craven AJ, Ashby MG, Pearson AJ: Regulation of prolactin receptor expression in ovine skin in relation to circulating prolactin and wool follicle growth status. J Endocrinol 2002, 172(3):605-614.
23. Cui, JX, Du HL, Liang Y, Deng XM, Li N, Zhang XQ: Association of polymorphisms in the promotor region of chicken prolactin with egg production. Poult Sci 2006, 85:26-31.
24. Juhn M, Harris PC: Molt of capon feathering with prolactin. Proc Soc Exp Biol Med 1958, 98(3):669-672.
25. Clevenger CV, Freier DO, Kline JB: Prolactin receptor signal transduction in cells of the immune system. J Enocrinol 1998, 157(2):187-197.
26. Craven AJ, Ormandy CJ, Robertson FG, Wilkins RJ, Kelly PA, Nixon AJ, Pearson AJ: Prolactin signaling influenced the timing mechanism of the hair follicle: analysis of hair growth cycles in prolactin receptor knockout mice. Endocrinology 2001, 142(b):2533-2539.
27. Foitzik K, Krause K, Nixon AJ, Ford CA, Ohnemus U, Pearson AJ, Paus R: Prolactin and its receptor are expressed in murine hair follicle epithelium, show hair cycle-dependent expression, and induce catagen. Am J Pathol 2003, 162(5):1611-1621.
28. Ostrowski LE, Andrews K, Potdar P, Matsuura H, Jetten A, Nettesheim P: Cloning and characterization of KPL2, a novel gene induced during ciliogenesis of tracheal epithelial cells. Am J Respir Cell Mol Biol 1999, 20(4):675-683.
29. Sironen A, Thomsen B, Andersson M, Ahola V, Vilkki J: An intronic insertion in KPL2 results in aberrant splicing and causes the immotile short-tail sperm defect in the pig. Proc Natl Acad Sci USA 2006, 103(13):5006-5011.
30. Primer3 website [http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi]
31. Ensembl Genome Browser [http://www.ensembl.org]
32. Sijben JW, Klasing KC, Schrama JW, Parmentier HK, van der Poel JJ, Savelkoul HF, Kaiser P: Early in vivo cytokine genes expression in chickens after challenge with Salmonella typhimurium lipopolysaccharide and

modulation by dietary n--3 polyunsaturated fatty acids. Dev Comp Immunol 2003, 27(6-7):611-619.

33. Staden Package Home Page [http://staden.sourceforge.net/]

4

Massive parallel sequencing of 12 genomes identifies protein affecting variants within QTL regions associated with the pulmonary hypertension syndrome in chicken.

M.G. Elferink[1], A.M. Closter[1], P. van As[2], D. Nikolic[1], H.J. Megens[1], H. Bovenhuis[1], R.P.M.A. Crooijmans[1], M.A.M. Groenen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University and Research Centre, Marijkeweg 40, PO Box 338, Wageningen, The Netherlands; [2]Hendrix Genetics Research, Technology & Services B.V., Spoorstraat 69, PO Box 114, Boxmeer, the Netherlands

Ready for submission

Abstract

The recent advances in massive parallel sequencing technologies have enabled rapid and cost-effective detection of all genetic variants within genomes. The detection of all genetic variants within a genome has further increased our ability to identify causative variants underlying quantitative trait loci (QTL). In this study, we combined a genome-wide association study with whole-genome resequencing to identify causative variants underlying the pulmonary hypertension syndrome (PHS) in chicken. PHS is a metabolic disease that has been linked to intense selection on growth rate and feed conversion ratio of modern broilers (meat-type chicken). PHS has become one of the most frequent causes of mortality within the broiler industry and leads to substantial economic losses and reduced animal welfare. In total, 18 QTL regions were identified in the genome-wide association study. In order to detect causative variants underlying these QTL regions, we sequenced the genomes of twelve individuals. To maximize the detection of causative variants we selected the individuals based on extreme phenotypes for PHS. In total 70 potential protein function affecting SNPs were detected in 28 genes within 13 out of the 18 QTL regions. Within 10 genes, at least one variant is predicted to affect the protein function and several genes have a clear functional relationship with PHS.

4.1 Introduction

The Pulmonary Hypertension Syndrome (PHS, also reffered to as Ascites Syndrome) is a metabolic disease that has been linked to intense selection on growth rate and feed conversion ratio of modern broilers (meat-type chicken) [1,2]. PHS has become one of the most frequent causes of mortality within the broiler industry and leads to substantial economic losses and reduced animal welfare [3,4]. Right ventricular failure as a result of pulmonary hypertension is the most frequent cause of PHS [5]. PHS resulting from right ventricular failure is characterized by a flaccid heart due to dilation and right ventricular hypertrophy, liver abnormalities, and ascites fluid excretion in the pericardium and abdominal cavity (Figure 4.1) [1,2,6,7,8]. The incidence of PHS can be influenced by factors that increase the oxygen demand, metabolic rate, heat production or resistance to blood flow in the lung [1,9,10,11]. Environmental conditions including altitude, temperature, lighting and ventilation can also contribute to increased PHS occurrence [2]. The main contributor to PHS is believed to be hypoxia that results from a disproportion between oxygen requirement and the cardiovascular ability to supply oxygen [7,12,13].

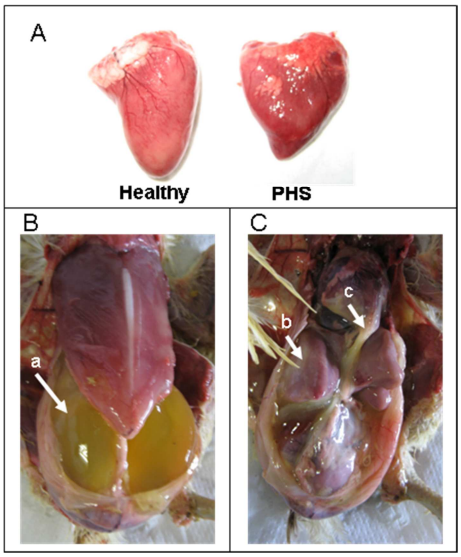


Figure 4.1 Clinical signs of PHS. A) A healthy chicken heart and a severely affected chicken heart due to PHS. Note that the PHS heart is flaccid and has an enlarged right ventricle. B) Dissected chicken suffering from PHS. The arrow (a) indicates ascites fluid in the abdominal cavity. C) Dissected chicken suffering from PHS after removal of ascites fluid and the ribcage. Arrow b indicates liver abnormality and arrow c indicates fluid in the pericardium cavity.

Despite intense research for several decades the molecular etiology of PHS remains unclear. Several QTLs have been identified in linkage analysis studies, thereby suggesting a polygenic and complex background for PHS [14]. Large confidence intervals of the identified QTLs resulted in a large number of potential candidate genes. Although for several genes the coding regions were sequenced using traditional Sanger sequencing, no causative variants were detected that contributed to PHS (Elferink unpublished results).

The development of high density genome-wide SNP assays has enabled genome-wide association studies (GWAS) [15]. Because GWAS does not need to be performed on a within family basis, which is the case for classical linkage analysis, large sample sizes can be used. These large sample sizes results in increased statistical power to identify [16] and map [15,17] QTLs, which is required for polygenic traits. GWAS has successfully identified QTLs involved in both monogenic and polygenic traits and diseases in human [18] and livestock species [19,20,21,22,23].

Recent developments in massive parallel sequencing (MPS) technologies [24,25,26,27] have enabled high-throughput identification of causative variants underlying a variety of traits and diseases [28,29,30,31,32,33]. Although currently too expensive for large sample sizes, costs can be reduced by careful selection of a small subset of individuals with *a priori* knowledge of the trait under investigation to maximize the detection of (rare) causative variants [34].

In this study, we performed a GWAS based on 895 animals genotyped with 17,790 SNPs and phenotypic observation from 8,158 offspring. In order to detect causative variants underlying the GWAS signals we re-sequenced the individual genomes of 12 of the 895 genotyped parents. To maximize the detection of variants involved in PHS, we selected 6 animals with extreme low and 6 animals with extreme high estimated breeding values (EBV) for RATIO, an indicator trait for PHS. In this paper we describe the first phase towards the detection of causative variants underlying PHS, in which we focus on the detection of potential protein affecting SNPs closely located near the most significant GWAS signals.

4.2 Material & Methods

GWAS

The animal population used, was a purebred broiler dam line originating from the White Plymouth Rock breed. The main selection criterion for this population was breast meat percentage. The effective population size of this population is approximately 100. The dataset for the GWAS consisted of 895 parents genotyped

with an Illumina Infinium iSelect Beadchip with 17,790 SNPs and 8,158 offspring that were phenotyped for the PHS indicator trait RATIO (right ventricular weight / total ventricular weight). RATIO is an indicator for the amount of right ventricular hypertrophy, and has been suggested as a good indicator trait for PHS [12,35,36].
All phenotyped chickens were kept under a cold temperature regime and increased CO_2 levels to challenge PHS development. The chickens were group housed with 20 birds/m2, had *ad libitum* access to commercial broiler feed containing 12,970 KJ/kg, and were exposed to 23h of light per day during the entire experiment. Except for the applied temperature schedule and increased CO_2 level, the chickens were kept under conditions that closely resemble commercial practice. All animals that died during the experiment were phenotyped on PHS traits. Surviving animals were sacrificed at 7 weeks of age and thereafter phenotyped for PHS related traits. The experiment was carried out by licensed and authorized personnel under approval of Hendrix Genetics, Boxmeer, the Netherlands.
ASReml software [37] was used to calculate associations between each SNP and indicator trait RATIO. For the analysis, only SNPs located on autosomal chromosomes were used. The following mixed model was used: $y_{ijk} = \mu + SNP_i + animal_j + e_{ijk}$, where

y_{ijk}= the average adjusted trait value of the 895 parents. The phenotypes of the offspring were adjusted for systematic environmental effects sex and the batch by stable interaction. The adjusted trait values were corrected for the contribution of the mate, averaged and assigned to a parent,

SNP_i= the fixed effect of the SNP genotype. A single SNP analysis was performed,

$animal_j$= the random genetic effect of Animal j. $Animal \sim N(0, A\sigma_a^2)$. A is the additive genetic relationship matrix accounting for all family relationships.

e_{ijk}= the random residual effect with $e \sim N(0, W\sigma_e^2)$. W is a diagonal matrix containing weights for each observation. The weight for each observation is the number of progeny of a parent.

The following model was used to calculate estimated breeding values:

$y_{jklmno} = \mu + sex_j + batch * stable_k + animal_o + e_{ijklmno}$, where

$y_{ijklmnop}$= the phenotype of individual ijklmnop, i.e. observations on 8,158 offspring,

sex_j,= effect of gender of the bird (j= 1,2, male or female),

$batch*stable_k$= effect of the interaction between batch and stable (k= 1, 2….,10). There were 5 batches and 2 stables,

$animal_j$= is the random genetic effect of Animal j. $Animal \sim N(0, A\sigma_a^2)$. A is the additive genetic relationship matrix accounting for all family relationships,

e_{ijk}= is random residual effect with $e \sim N(0, I\sigma_e^2)$.

Estimated breeding values for the 895 parents were selected from the file containing solutions for all animals in the pedigree file.

Table 4.1 Estimated breeding values and alignment statistics for the selected animals.

Animal ID	Sire ID	Dam ID	# Off [1]	EBV RATIO	Sequence coverage[2] (fold)	Assembly coverage[3] (%)	Assembly coverage 4X[4] (%)
9668	1966	2403	145	-6.4	12.5	91.9	88.1
9259	3742	4697	48	-6.1	7.7	91.5	81.6
8699	9538	8915	41	-5.8	11.9	92.0	88.6
9660	1242	1097	146	-5.4	11.9	92.0	88.3
9439	5140	4957	73	-5.1	11.4	92.2	88.3
9653	1229	1873	33	-5.1	10.9	92.0	88.0
9841	9614	9636	56	3.9	13.9	92.1	88.9
9801	9056	9583	112	4.0	11.5	91.9	87.8
8993	6459	6654	56	4.5	16.1	92.3	89.9
8711	9538	9854	156	4.7	14.2	92.2	89.3
9869	1831	8516	179	5.7	14.4	92.7	89.0
8788	901	8947	38	7.8	12.2	92.3	88.7

[1] # off = number of phenotyped offspring on which the EBV for RATIO was calculated. [2] The average sequence depth of each base in the reference genome that is covered by at least 1 read. The used reference genome without chrUn_random and the artificial centromeres is 1,016,609,635 bp. [3] Percentage of the reference genome that is covered by at least one read. [4] Percentage of the reference genome that is covered by at least 4 reads.

Whole-genome resequencing and SNP discovery

In total, twelve animals were selected for whole-genome resequencing based on their breeding values for RATIO (Table 4.1). Six animals had extreme low breeding values and six animals had extreme high breeding values. Animals within one of the extreme phenotype groups were not allowed to be full- or half-sibs. Due to our experimental design, sires had much more offspring compared to dams. As breeding values for animals with a low number of offspring are regressed towards the mean, extreme animals, i.e. the animals selected for resequencing, were all male. DNA was extracted from whole-blood using the Gentra DNA extraction kit (Qiagen). The DNA was randomly sheared and libraries were made from fragments with an average length of 200 bp. Each animal was sequenced (100 bp, paired-end) in one lane of a flow cell on the Illumina HiSeq 2000 platform. The workflow for SNP discovery was as follows. Custom made python scripts were used to trim the read on base quality. If three bases in row had a base quality of less than 20, the read was trimmed from the first base that was below this threshold. Both mates of the paired-end reads were required to be at least 36 base pairs in size. MOSAIK assembler software [38] was used to align the paired-end reads to the reference genome of the chicken (build WASHUC2). Except for chromosome unassigned, all known chromosomes and linkage groups were used (total length 1,016,609,635 bp). Alignment parameters were as follows: hash size= 15, maximum percentage of mismatches allowed= 7%, alignment candidate threshold= 30, maximum hash position= 100. The option 'aligned read length to count mismatches' (-mmal) was applied, and the threshold for the minimum percentage of read length that needs to be aligned (-minp) was set to 50%. To increase the alignment the banded Smith-Waterman algorithm was used (bw= 41) and the reference genome was converted to a jump database. Alignment files were sorted using MosaikSort with the 'allow all fragments lengths evaluating unique read pairs' option.

The mpileup function of SamTools version 0.1.12a [39] was used to call variants for all twelve animals simultaneous. The view option of bcftools [39]) was used to call the genotype at each variant for each animal. Custom made python scripts were used to remove genotype calls per animal with a genotype quality less than 20, with a base coverage less than 4 or more than 30, and for tri-allelic SNPs. In this study, we focused on SNPs only and, therefore, removed all genotype call for indels. Each SNP position where at least 4 animals had a genotype call that passed filter criteria, was regarded as a putative SNP.

In order to estimate the SNP false discovery rate (FDR), we focused on a large region in the genome in which all animals were clearly homozygous for at least one haplotype. We regarded each heterozygous SNP call within this homozygous region

to be a false positive. For each animal, the FDR was calculated by dividing the number of heterozygous SNP within this homozygous region by the number of bases that were sufficiently covered for genotype calling (4 - 30 fold coverage). In order to estimate the false negative rate (FNR), we compered the genotype data for the 18k SNP assay with the genotype call from MPS, and for each animal separately. More specifically, a SNP was considered a false negative if a heterozygous call in the 18k assay was called homozygous within the MPS data.

Functional annotation of SNPs

The gene-based analysis of ANNOVAR software [40] was used to functionally annotate the putative SNPs. For each putative SNP, the position (exonic, intronic, intergenic, 5'UTR, 3'UTR, splice acceptor or donor site, downstream, or upstream) and the functional annotation (nonsynonymous, synonymous, stop codon gain, stop codon lost) were determined based on the reference genome (WASHUC2) and gene annotation from Ensembl [41]. Standard setting for gene-based analysis of ANNOVAR were used.

SIFT software version 4.0.3 [42] was used to determine the effect of nonsynonymous variants on protein function. SIFT software predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. Standard setting for SIFT predictions were used – a prediction score of less than 0.05 was regarded to affect protein function - and sequence prediction were based on the NCBI nonredundant database (2008).

4.3 Results

GWAS

In total, 9,653 autosomal SNPs (chromosome 1-28) segregating within the population were used for the single SNP association analysis (Figure 4.2). We decided to focus on the most significant SNPs within our analysis and choose a –log10(p-value) threshold of 3 which corresponds to a FDR ≤ 0.38. In total, 25 SNPs passed this threshold (Table 4.2). Three of these SNPs exceeded a –log10(p-value) of 4 which corresponds to a FDR ≤ 0.24. In order to detect potential causative variants, we decided to focus on a window of 200kb surrounding each of the 25 SNPs (100kb upstream and downstream of each SNP). However, if a gene was partially located within the window, the window size was increased in order to include the entire gene. Overlapping windows were merged, which resulted in 18 different chromosomal regions, hereafter referred to as QTL regions (Table 4.2). Out of the 18 QTL regions, 14 regions contained a single SNP with high significance and 4 regions contained 2 to 4 SNPs with high significance (Table 4.2).

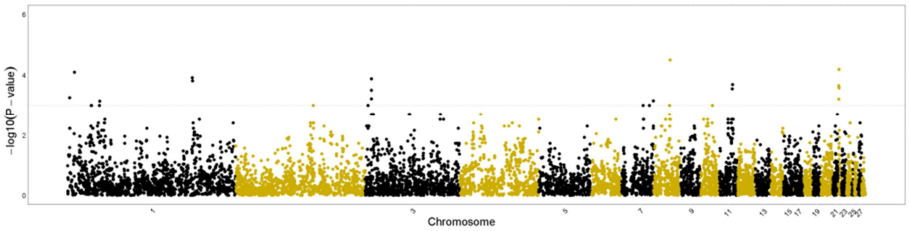


Figure 4.2 Genome-wide plot of all SNPs used in the association study. On the y-axis the –log10(p-value) is given. On the x-axis the chromosome location is given (GGA1-28, only uneven chromosomes are labeled). The dashed grey line indicates the –log10(p-value) threshold of 3.

Whole-genome resequencing and SNP discovery

DNA of each individual animal was sequenced in a single lane of a flow cell resulting in 7.8 - 16.4 Gb of sequence per animal. After quality trimming and alignment of the short reads, the percentage of bases in the reference genome covered by at least one read varied between 91.5% and 92.7% for the 12 animals sequenced (Table 4.1). The percentage of bases in the reference genome covered by at least four reads, and therefore sufficient for genotype calling, varied between 81.6% and 89.9% (Table 4.1). The average sequence depth for each covered base in the reference genome varied between 7.1 and 14.9 (Table 4.1).

In total, 7.62 million putative SNPs were detected compared to the reference genome (Table 4.3). Of these SNPs, 6.17 million SNPs were segregating within the twelve animals. On average, 4.56 million SNPs were detected in each individual compared to the reference genome (that is, the individual is either heterozygous or homozygous for the non-reference allele) (Table 4.3). Within each animal, on average, two million heterozygous SNPs are detected (ranging from 1.69-2.58 million). The SNP FDR was estimated based on the number of heterozygote SNP calls within a 4.73 Mb large regions on chromosome 1 (129,085,000-133,815,000 bp), for which all animals clearly are homozygous for a single haplotype (Figure 4.3). Based on this region, we estimate the SNP FDR to be 0.0013% per nucleotide position in the reference genome (ranging from 0.0009-0.0018%). This indicates that, on average 12,794 false positive heterozygous SNPs will be detected per individual in the entire genome. Based on the heterozygous SNP calls from the 18k SNP assay, we estimate the SNP FNR to be 12.3% per heterozygous genotype call within entire genome (ranging from 6.5-24.5%). This indicates that, on average, 282,937 heterozygous SNPs are missed per individual (ranging from 157,672-548,131). In total, corrected for the FDR, and the FNR, each individual contains, on average, 2.27 million heterozygous SNPs in the entire genome (ranging from 1.86-

3.03 million). The average heterozygous SNP frequency per individual is 2.24 SNPs kb^{-1} (ranging from 1.83-2.98 SNPs kb^{-1}).

Table 4.2 The most significant SNPs detected in the GWAS.

SNP ID	Chr	Position	Region (+/- 100kb)[1]		-log10(p-value)	FDR
rs14789557	1	2,124,738	2,024,738	2,224,738	3.26	0.38
rs16079719	1	7,880,128	7,780,128	7,980,128	4.10	0.24
rs13841399	1	27,984,448	27,884,448	28,084,448	3.00	0.38
rs13747646	1	28,523,801	28,423,801	28,632,749	3.00	0.38
rs14811108	1	37,855,143	37,725,119	37,955,143	3.00	0.38
rs15236245	1	38,231,869	38,117,669	38,331,869	3.15	0.38
rs14899763	1	148,951,482	148,754,342	149,273,569	3.91	0.24
rs13952858	1	149,173,569			3.81	0.24
rs14218633	2	92,844,355	92,738,644	92,981,551	3.00	0.38
rs15257935	3	3,581,789	3,462,639	3,742,795	3.00	0.38
rs16225894	3	7,578,369			3.88	0.24
rs15272751	3	7,646,096	7,478,369	7,930,821	3.51	0.27
rs14317011	3	7,804,975			3.22	0.38
rs14617579	7	25,555,092	25,228,473	25,655,092	3.00	0.38
rs13599609	7	33,065,420	32,965,420	33,165,420	3.00	0.38
rs16615527	7	37,568,883	37,468,883	37,668,883	3.15	0.38
rs15920819	8	19,019,021	18,919,021	19,119,021	3.00	0.38
rs15921649	8	19,584,685	19,484,685	19,684,685	4.50	0.24
rs15580567	10	13,992,236	13,885,340	14,092,236	3.00	0.38
rs14965732	11	15,491,746	15,431,743	15,946,571	3.55	0.27
rs14965814	11	15,744,074			3.69	0.26
rs15187369	22	531,582			3.66	0.26
rs16183544	22	557,110	421,583	1,006,582	3.21	0.38
rs16183608	22	786,138			4.19	0.24
rs15187555	22	906,582			3.58	0.27

[1] The region of interest was determined +/- 100kb of the position of the SNP, unless genes were partially located at one of the ends. In that case, the window size was increased in order to include the entire gene. FDR= false discovery rate.

Table 4.3 Number of SNPs detected.

	Reference total[1]	Reference Individual[2]	Segregating total[3]
Nonsynonymous	23,568	13,193	18,492
Synonymous	59,947	34,266	48,822
Stop gain	275	137	222
Stop lost	19	10	12
Splice acceptor or donor site	761	440	567
5'UTR/3'UTR	32,398	18,798	26,468
downstream/upstream	178,299	99,995	144,668
Intronic	3,028,208	1,794,167	2,487,645
Intergenic	4,301,335	2,601,017	3,446,460
ncRNA	613	321	489
Total	7,625,423	4,562,343	6,173,845

[1] The total number of SNPs detected compared to the reference genome in which the non-reference allele is detected in at least 1 of the twelve animals. [2] The average number of SNPs detected in each animal compared to the reference genome (the individual is either heterozygous or homozygous for the non-reference allele). [3] The total number of SNPs detected segregating within the 12 animals.

Functional annotation of SNPs

Within the 18 QTL regions a total of 37,024 SNPs are detected that segregated within the twelve animals. Of these SNPs, 340 are located in coding regions and 4 are located in splice acceptor or donor sites. The 340 coding SNPs include 64 nonsynonymous, 2 premature stops, and 274 synonymous SNPs. The nonsynonymous, stop codon affecting, and splice acceptor or donor site SNPs are located within 28 different genes in 13 of the 18 QTL regions (Table 4.4). Of the 64 nonsynonymous SNPs (NS-SNPs) that were submitted to SIFT, 14 are predicted to affect protein function (6 with low confidence), and 46 are predicted to have no effect on protein function. For 4 NS-SNPs no prediction could be made as there are no orthologs found in the database. One of the NS-SNPs is highly significant in the GWAS (rs13952858, chr1 149,173,569 bp, -log10(p-value)= 3.81). This SNP is located in *DOCK9* and predicted not to affect protein function. The 14 SNPs that are predicted to affect protein function are located in 8 genes (Table 4.4). Four genes contain a single SNP that is predicted to affect protein function (*PTPRB, CAPN13, Q5ZHW0, KIAA1199*), and four genes have two or three of these SNPs (*DMTF1*= 2, *KIZ*= 3, *GTF2A1L*= 2, and *CAPN14*= 3). The two premature stops are detected in two different genes (*ZNF236* and *LOC771515*).

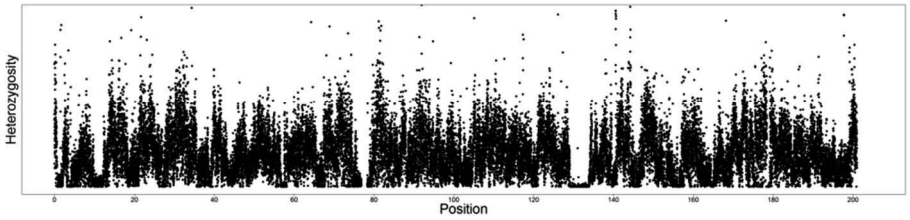


Figure 4.3 Heterozygosity across chromosome 1 for the 12 individuals. On the x-axis the chromosomal position is shown in Mb. On the y-axis the heterozygosity is given as a percentage of heterozygous SNP within a 5000 kb window, correct for the number of bases covered. Note the clear homozygous region at 129-133 Mb.

The allele frequency difference between the low and high EBV groups is, in general, not very high for all 16 variants that are predicted to affect protein function (14 NS-SNPs and the 2 premature stops). A notable exception is the premature stop in *ZNF236* in which an allele frequency difference of 33% is found between the two groups. This variant is, moreover, heterozygous in all high EBV individuals.
For six genes, the allele that is predicted to affect protein function is exclusively found in one of the two EBV groups. Seven variants in 5 genes are detected exclusively in the low EBV group; *DMTF1* (7,930,486 bp), *KIZ* (all three variants), *GTF2A1L* (7,569,531 bp), *CAPN14* (7,652,889 bp), and *KIAA119* (13,945,539 bp). All three variants in *KIZ* are detected within a single individual (ID= 9668). Two variants located 2 different genes are detected exclusively in the high EBV group; *CAPN14* (7,662,958 bp), and the premature stop in *LOC771515*.

4.4 Discussion

In this study, we combined the power of genome-wide association studies with the high-throughput capacity of massive parallel sequencing to detect potential variants involved in PHS, a polygenic trait in chicken. In the first phase to achieve this goal, we focused on possible protein affecting SNPs located near the most significant GWAS signals.
The SNP genotypes obtained by the whole-genome sequencing efforts are of high quality and cover nearly the complete reference genome. The high quality of the SNP genotypes is reflected by the low FDR. The average heterozygous SNP frequency per individual of 2.24 SNPs kb^{-1} is substantially lower than the 4.28 SNPs kb^{-1} previously detected within a single broiler animal [43]. This difference is likely caused by the effective population size of the broiler line used in that study (n= 800) and the population used in our study (n= 100). Due to the small effective population size it is expected that the nucleotide diversity is lower.

Table 4.4 Predicted effect of the amino acid change of the coding SNPs within the QTL regions.

Chr	Position	SNP	Gene	AA change[1]	SIFT Prediction[2]		Genotype[3]	DIF[4]
1	7,847,975	T/C	*LOC776661*	L>S	TOL	0.23	(3,2,0),(0,4,1)	0.40
1	7,879,238	C/T	*DCR1C*	R>K	TOL	0.21	(3,1,1),(2,3,0)	0.00
1	7,930,486	C/T	*DMTF1*	E>K	AFF*	0.03*	(4,2,0),(6,0,0)	0.17
1	7,930,888	C/T	*DMTF1*	C>Y	AFF*	0.05*	(0,2,1),(0,3,0)	0.17
1	28,584,789	T/C	*CG060*	SS	ND	ND	(3,2,0),(4,1,0)	0.10
1	28,614,315	A/G	*TMEM168*	I>V	TOL	0.33	(2,4,0),(2,2,0)	0.08
1	28,614,471	G/A	*TMEM168*	V>I	TOL	1.00	(0,3,1),(2,2,1)	0.23
1	28,614,567	A/G	*TMEM168*	I>V	TOL	1.00	(2,4,0),(3,1,1)	0.03
1	28,632,203	A/G	*TMEM168*	SS	ND	ND	(1,0,0),(1,2,0)	0.33
1	28,632,375	A/G	*TMEM168*	I>V	TOL	1.00	(2,1,0),(1,2,1)	0.33
1	37,746,299	C/T	*PTPRB*	R>Q	TOL	0.68	(4,2,0),(3,1,1)	0.13
1	37,747,443	C/T	*PTPRB*	R>Q	TOL	0.38	(4,2,0),(3,1,1)	0.13
1	37,755,710	G/A	*PTPRB*	L>F	TOL	0.06	(1,2,0),(2,2,0)	0.08
1	37,758,309	C/T	*PTPRB*	A>T	AFF	0.01	(4,2,0),(4,1,1)	0.08
1	37,759,225	A/G	*PTPRB*	I>T	TOL	1.00	(4,1,0),(4,1,1)	0.15
1	38,196,684	A/G	*LGR5*	N>S	TOL	0.28	(5,1,0),(6,0,0)	0.08
1	38,229,608	G/A	*CCDC131*	T>I	TOL	0.22	(5,1,0),(6,0,0)	0.08
1	38,231,432	C/T	*CCDC131*	S>N	TOL	0.68	(2,3,0),(3,1,1)	0.00
1	148,802,881	A/G	*EBI2*	T>A	TOL	0.11	(4,2,0),(2,3,0)	0.13
1	148,947,512	G/A	*DOCK9*	V>I	TOL	0.37	(4,1,0),(6,0,0)	0.10
1	148,947,924	A/G	*DOCK9*	N>S	TOL	0.15	(6,0,0),(4,2,0)	0.17
1	148,951,482**	T/G	*DOCK9*	F>L	TOL	1.00	(0,0,6),(1,3,2)	0.42
1	148,952,612	A/G	*DOCK9*	I>V	TOL	0.64	(2,3,1),(1,2,1)	0.08
1	149,022,744	C/T	*Q90WH8*	P>L	TOL	0.07	(1,4,0),(4,0,0)	0.40
1	149,039,177	G/T	*Q90WH8*	E>D	TOL	0.42	(3,0,0),(0,2,1)	0.67
1	149,040,341	C/T	*Q90WH8*	S>L	TOL	0.81	(4,1,0),(5,1,0)	0.02
1	149,163,240	A/G	*FARP1*	SS	ND	ND	(2,3,0),(5,1,0)	0.22
1	149,167,822	T/A	*FARP1*	E>D	TOL	1.00	(4,1,0),(6,0,0)	0.10
2	92,922,982	G/A	*ZNF236*	R>X	SC	ND	(4,2,0),(0,6,0)	0.33
2	92,946,985	C/A	*ZNF236*	E>D	TOL	0.12	(0,2,4),(0,6,0)	0.33
2	92,955,528	T/C	*ZNF236*	N>S	TOL	0.48	(0,5,0),(3,3,0)	0.25
2	92,962,465	T/C	*ZNF236*	K>R	TOL	0.16	(3,1,0),(0,6,0)	0.38
2	92,962,565	C/T	*ZNF236*	A>T	TOL	0.42	(1,3,1),(0,6,0)	0.00
2	92,981,520	C/T	*ZNF236*	R>H	TOL	0.14	(0,4,0),(1,5,0)	0.08
3	3,542,561	G/A	*KIZ*	P>L	AFF	0.01	(4,1,0),(6,0,0)	0.10
3	3,544,499	C/T	*KIZ*	R>H	AFF	0.01	(5,1,0),(5,0,0)	0.08
3	3,544,604	C/T	*KIZ*	R>H	AFF	0.01	(5,1,0),(6,0,0)	0.08
3	3,636,817	A/G	*C20orf74*	I>V	TOL	1.00	(4,1,0),(6,0,0)	0.10
3	3,686,313	G/A	*C20orf74*	V>I	TOL	0.52	(2,3,0),(5,1,0)	0.22

Table 4.4 continued…

Chr	Position	SNP	Gene	AA change[1]	SIFT Prediction[2]		Genotype [3]	DIF [4]
3	7,568,423	A/G	*GTF2A1L*	S>P	AFF	0.03	(0,3,3),(0,3,2)	0.05
3	7,569,191	T/G	*GTF2A1L*	N>H	TOL	0.10	(0,3,1),(1,3,2)	0.04
3	7,569,531	G/T	*GTF2A1L*	S>R	AFF*	0.01*	(5,1,0),(5,0,0)	0.08
3	7,652,889	C/T	*CAPN14*	P>L	AFF	0.00	(5,1,0),(5,0,0)	0.08
3	7,662,958	A/G	*CAPN14*	N>S	AFF	0.00	(2,0,0),(1,1,0)	0.25
3	7,664,707	A/C	*CAPN14*	Q>P	AFF*	0.02*	(0,3,2),(1,1,3)	0.00
3	7,696,520	A/G	*EN23741***	Y>C	NO	ND	(4,1,0),(6,0,0)	0.10
3	7,772,411	C/T	*CAPN13*	P>L	AFF*	0.00*	(3,3,0),(3,2,1)	0.08
3	7,785,156	G/A	*CAPN13*	S>N	TOL	0.18	(0,2,1),(0,2,1)	0.00
3	7,787,010	A/T	*CAPN13*	H>L	TOL	0.29	(4,1,0),(3,3,0)	0.15
3	7,792,019	G/A	*CAPN13*	V>M	TOL	0.05	(5,1,0),(6,0,0)	0.08
7	33,008,112	G/A	*SPOPL*	G>E	TOL	0.58	(5,1,0),(5,1,0)	0.00
7	33,106,861	C/T	*LOC771515*	A>T	NO	ND	(4,0,0),(2,1,0)	0.17
7	33,106,873	G/A	*LOC771515*	Q>X	SC	ND	(4,0,0),(2,2,0)	0.25
7	33,106,894	C/T	*LOC771515*	A>T	NO	ND	(2,1,0),(3,1,0)	0.04
7	33,106,920	C/G	*LOC771515*	W>S	NO	ND	(0,1,3),(0,1,3)	0.00
8	19,573,784	T/A	*DNAJB4*	Y>F	TOL	0.35	(0,2,4),(0,5,1)	0.25
8	19,600,476	C/T	*Q5ZHW0*	R>C	AFF*	0.01*	(4,2,0),(2,4,0)	0.17
8	19,629,146	C/T	*NEXN*	S>N	TOL	0.27	(6,0,0),(5,1,0)	0.08
8	19,629,305	A/T	*NEXN*	M>K	TOL	0.70	(6,0,0),(5,1,0)	0.08
10	13,945,539	T/C	*KIAA1199*	I>M	AFF	0.01	(1,3,0),(5,0,0)	0.38
10	13,947,792	T/C	*KIAA1199*	I>V	TOL	0.32	(1,3,1),(5,0,1)	0.33
10	13,977,136	C/T	*KIAA1199*	R>Q	TOL	0.44	(5,1,0),(3,2,0)	0.12
10	13,985,305	T/C	*KIAA1199*	SS	ND	ND	(1,3,0),(4,0,1)	0.18
10	13,985,916	C/G	*KIAA1199*	A>P	TOL	0.37	(1,4,1),(1,3,1)	0.00
10	13,987,625	T/C	*KIAA1199*	E>G	TOL	0.38	(4,1,0),(3,3,0)	0.15
10	14,001,168	C/T	*TMEM2*	V>I	TOL	0.39	(0,2,1),(1,2,2)	0.07
10	14,021,565	T/G	*TMEM2*	K>Q	TOL	0.19	(1,3,1),(1,3,1)	0.00
10	14,038,198	T/C	*TMEM2*	I>V	TOL	1.00	(5,1,0),(3,3,0)	0.17
11	15,558,407	A/G	*WWOX*	I>V	TOL	1.00	(0,4,2),(0,4,0)	0.17
22	851,431	C/T	*DOCK5*	A>T	TOL	0.38	(1,3,0),(0,6,0)	0.13

[1] AA change= amino acid change. [2] SIFT predictions of AA change. Values below 0.05 are predicted to affect protein function and above 0.05 are tolerated. [3] Genotype of the SNP in either low EBV (brackets left) or high EBV (brackets right). For both, the left value indicates the number of animals homozygous for the reference allele, the middle indicates the number of heterozygous animals, and the right value indicates the number of animals homozygous for the non-reference allele. [4] DIF= allele frequency difference between the low and high EBV phenotype groups. * prediction with low confidence due to either a low number of orthologous sequences or due to low sequence diversity of the orthologous sequences. ** SNP rs13952858, -log10(p-value)= 3.81 in the GWAS. *** abbreviation for ENSGALG00000023741. TOL= tolerated. AFF= affected. SC= stop codon. NO= no ortholog. ND= not determined. SNP= reference allele/ non-reference allele. EBV= estimated breeding value. Grey shaded areas correspond to chromosomal regions at significant GWAS SNPs. SS= splice acceptor or donor site.

Within the GWAS we identified 18 QTL regions associated with PHS. These results suggest that PHS is a polygenic trait influenced by a large number of loci each with a small phenotypic effect (Figure 4.2, Table 4.2). This observation is in agreement with a previous linkage mapping study [14], although other studies have proposed that major genes are involved in PHS [44,45,46]. However, these studies were based on different broiler populations, and it is possible that other alleles – possibly with large phenotypic effects - are segregating in those populations. Moreover, we cannot exclude the presence of additional loci, either with small or large phenotypic effect, located on chromosomes that were not included in the association analysis such as chromosome Z, and the microchromosomes that are absent from the reference genome [47,48,49]. Furthermore, loci might have been undetected due to missing or insufficient linkage disequilibrium with assayed markers. Based on LD measurements it has been suggested that genome-wide assays in broiler populations need to include at least 100k SNPs to cover all haplotypes within the genome [50]. The SNP assay used in this study existed of 18k SNPs and, therefore, it is likely that some regions in the genome will not be sufficiently covered.

There are three QTL regions in our study that overlap, or are closely located to, previously identified QTL regions [14]. Although two QTL regions in our study do not physically overlap with the previously identified QTL regions, it has to be mentioned that the confidence intervals in the linkage mapping study of Rabie *et al.* (2005) are large. Therefore, the QTL regions identified in both studies might be associated with the same underlying causative variant(s). The QTL region on chromosome 2 (92.8 Mb) is near significant QTL regions for right ventricular weight as percentage of body weight, total ventricular weight as percentage of body weight, and RATIO. The QTL identified on chromosome 8 (19.0 Mb) is near significant QTL regions for total ventricular weight as percentage of body weight, and body weight at 5 weeks under PHS inducing conditions. The third QTL region, located on chromosome 10 (13.9 Mb), co-localizes with significant QTLs for the trait mortality due to PHS, and body weight at 5 weeks under PHS inducing conditions.

To detect causative variants that underlie the QTL regions identified in our GWAS, we decided to re-sequence the genomes of 12 individuals. To maximize the detection of variants involved in PHS, we decided to select 6 animals at both extremes of the estimated breeding value distribution. In this study we decided to focus on SNPs that could affect protein function - such as nonsynonymous, splice acceptor or donor site, and stop codons – located within 100kb up- or downstream of the GWAS signals. We are aware that other sources of variants, such as small insertion/deletions (indels) and large structural variations (insertions, deletions,

duplications, inversions and translocations) may also contribute to phenotypic variation by affecting coding regions or even entire genes. Moreover, synonymous variants in genes can result in changes in gene expression due to changed efficiency of translation (referred to as codon bias [51]) and intronic variants can alter gene expression [52] or result in alternative splicing [53]. Variants located in intergenic regions, such as promoter, enhancer, and silencer regions, can also result in altered gene expression. In addition, most variants involved in complex diseases -in humans- are involved in gene regulation and are not located within coding regions [54]. Furthermore, true causative variants - especially rare variants - can be located several megabases away from GWAS signals [55]. Nevertheless, in the current study, we decided to first focus on SNPs that affect protein function as they are relatively easy to detect with existing tools. Moreover, we decided to use a relatively small window size as we assume that most causative variants are located near the original GWAS signals.

In total, 28 genes contained at least one NS-SNP, premature stop, or splice acceptor or donor site SNPs (Table 4.4). Within 10 genes at least one variant is predicted to affect protein function. The known biological functions of two genes - *PTPRB* and *ZNF236* - could be directly related to PHS. The gene encoding protein-tyrosine phosphatase receptor-type beta (*PTPRB*) is expressed in endothelium, developing outflow tract of the heart, and developing heart valves in mice [56]. *PTPRB* is, moreover, essential for cardiovascular development [56]. Because the aetiology of PHS might be traced back as far as the embryonic stage [57], genes involved in cardiac development are obvious candidates. In addition, *PTPRB* activity is essential for maintenance and remodelling of blood vessels in mice [58]. Vascular remodelling is a well-known mechanism involved in hypertension [59,60]. The SNP variant that is predicted to affect protein function is an alanine to threonine substitution in amino acid 362 (protein ID= ENSGALP00000016311). This amino acid substitution is located within one of the 16 fibronectin type 3 domains located within the protein. These domains are part of the extracellular receptor-like domain of the protein.

ZNF236 is a Kruppel-like zinc-finger gene that is upregulated in human mesangial cells in response to elevated levels of d-glucose [61]. Mesangial cells are smooth muscle cells around blood vessels in the kidney that regulate blood flow though the capillaries. Although knowledge on the true function of *ZNF236* is limited, the regulation by d-glucose and its expression in cells that regulate blood flow are interesting with respect to PHS. The premature stop affects only the very last amino acid of the protein (arginine>stop in amino acid 1848, protein ID= ENSGALP00000022167), which could indicate that the premature stop will likely

have a limited influence on the protein structure and function. Nevertheless, we did not observe animals that were homozygous for this premature stop, thereby suggesting that the homozygous state might be lethal and thus affects protein function. If a variant is lethal, deviation from Hardy-Weinberg equilibrium (HWE) is expected. However, because the number of animals sequenced is small, the absence of homozygotes for these variants appears not to significantly deviate from HWE (p-value *ZNF236*= 0.34, calculation done with Haploview software [62,63]). Noticeably, all six animals in the high EBV group (PHS susceptible) are heterozygous for the premature stop in *ZNF236*, while only two (out of six) are heterozygous in the low EBV groups. This observation suggests that the premature stop in *ZNF236* might contribute to PHS susceptibility. However, large sample sizes are needed to confirm this observation.

Although 18 genes did not contain variants that are predicted to affect protein function, two of them - *NEXN* and *WWOX* – have biological functions that can be directly related to PHS. Variants in the gene encoding Nexilin (*NEXN*) are known to cause dilated and hypertrophic cardiomyopathy in human [64,65]. These cardiovascular diseases show similar characteristics to PHS in chicken, such as hypertrophy of the heart and congestive heart failure [66]. A study on the gene encoding WW domain-containing oxidoreductase (*WWOX*) shows a contribution of *WWOX* to aerobic metabolism and the regulation of reactive oxygen species [67]. Reactive oxygen species have been suggested to be involved in the aetiology of PHS [68]. The involvement of *WWOX* to aerobic metabolism could be linked to contribution of oxygen demand and metabolic rate that are involved in PHS development [1,9,12]. Both *NEXN* and *WWOX* did not contain variants that are predicted to affect protein function. The coding variants that are detected in these genes did not have a large allele frequency difference between the two EBV groups, although both variants in *NEXN* were exclusively detected within the high EBV group. Although *NEXN* has an acceptable coverage throughout the coding regions of the gene, exons 1 and 9 of *WWOX* are poorly or not covered by sequence reads. For WWOX additional sequencing efforts will be needed to detect all SNPs within this gene. Furthermore, both genes are obvious targets for in depth studies on regulatory variants and gene expression.

In conclusion, we combined the power of genome-wide association with high-throughput capacity of massive parallel sequencing to detect 70 potential protein function affecting SNPs that might affect susceptibility or resistance to PHS. Within 10 genes, at least one variant is predicted to affect the protein function.

Acknowledgments

The authors would like to thank Tineke Veenendaal for performing the genotyping experiments. This study was part of “The characterisation of genes involved in pulmonary hypertension syndrome in chicken” project funded by the Dutch Technology Foundation (STW). Project number 07106.

References

1. Decuypere E, Buyse J, Buys N (2000) Ascites in broiler chickens: exogenous and endogenous structural and functional causal factors. World's Poultry Science Journal 56: 367-377.
2. Balog JM (2003) Ascites Syndrome (Pulmonary Hypertension Syndrome) in Broiler Chickens: Are We Seeing the Light at the End of the Tunnel? Avian and Poultry Biology Reviews 14: 99 -126.
3. Maxwell MH, Robertson GW (1998) UK survey of broiler ascites and sudden death syndromes in 1993. British Poultry Science 39: 203 - 215.
4. Pavlidis HO, Balog JM, Stamps LK, Hughes JD, Jr., Huff WE, *et al.* (2007) Divergent Selection for Ascites Incidence in Chickens. Poult Sci 86: 2517-2529.
5. Julian RJ (2005) Production and growth related disorders and other metabolic diseases of poultry - A review. The Veterinary Journal 169: 350-369.
6. Olkowski AA, Korver D, Rathgeber B, Classen HL (1999) Cardiac index, oxygen delivery, and tissue oxygen extraction in slow and fast growing chickens, and in chickens with heart failure and ascites: a comparative study. Avian Pathology 28: 137 - 146.
7. Currie RJW (1999) Ascites in poultry: recent investigations. Avian Pathology 28: 313 - 326.
8. Baghbanzadeh A, Decuypere E (2008) Ascites syndrome in broilers: physiological and nutritional perspectives. Avian Pathology 37: 117 - 126.
9. Scheele CW, Decuypere E, Vereijken PFG, Schreurs FJG (1992) Ascites In Broilers. 2. Disturbances In The Hormonal-Regulation Of Metabolic-Rate And Fat-Metabolism. Poultry Science 71: 1971-1984.
10. Wideman R, Kirby Y (1995) A pulmonary artery clamp model for inducing pulmonary hypertension syndrome (ascites) in broilers. Poultry Science: May: 805-812.
11. Julian RJ (2000) Physiological, management and environmental triggers of the ascites syndrome: a review. Avian Pathology 29: 519 - 527.

12. Julian RJ (1993) Ascites in poultry. Avian Pathology 22: 419 - 454.
13. Decuypere E, Hassanzadeh M, Buys N, Buyse J (2005) Further insights into the susceptibility of broilers to ascites. The Veterinary Journal 169: 319-320.
14. Rabie T, Crooijmans R, Bovenhuis H, Vereijken A, Veenendaal T, *et al.* (2005) Genetic mapping of quantitative trait loci affecting susceptibility in chicken to develop pulmonary hypertension syndrome. Animal Genetics 36: 468 - 476.
15. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6: 95-108.
16. Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405: 847-856.
17. Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat. Rev. Genet. 2: 91-99.
18. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences 106: 9362-9367.
19. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, *et al.* (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat. Genet. 40: 449-454.
20. Karlsson EK (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. Nature Genet. 39: 1321-1328.
21. Duijvesteijn N, Knol E, Merks J, Crooijmans R, Groenen M, *et al.* (2010) A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC Genetics 11: 42.
22. Mai MD, Sahana G, Christiansen FB, Guldbrandtsen B (2010) A genome-wide association study for milk production traits in Danish Jersey cattle using a 50K single nucleotide polymorphism chip. J Anim Sci 88: 3522-3528.
23. Orr N, Back W, Gu J, Leegwater P, Govindarajan P, *et al.* (2010) Genome-wide SNP association–based localization of a dwarfism gene in Friesian dwarf horses. Animal Genetics 41: 2-7.
24. Mardis ER (2008) Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics 9: 387-402.
25. Metzker ML (2010) Sequencing technologies - the next generation. Nat. Rev. Genet.: 11: 31-46.

26. Eid J, Fehr A, Gray J, Luong K, Lyle J, *et al.* (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323: 133-138.
27. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, *et al.* (2010) Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 327: 78-81.
28. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature: 461: 272-276.
29. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. 42: 30-35.
30. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences 106: 19096-19101.
31. Sobreira NLM, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, *et al.* (2010) Whole-Genome Sequencing of a Single Proband Together with Linkage Analysis Identifies a Mendelian Disease Gene. PLoS Genet 6: e1000991.
32. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, *et al.* (2010) Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science 328: 636-639.
33. Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, *et al.* (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature 464: 1039-1042.
34. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. 11: 415-425.
35. Pakdel A, Van Arendonk JAM, Vereijken ALJ, Bovenhuis H (2005) Genetic parameters of ascites-related traits in broilers: effect of cold and normal temperature conditions. British Poultry Science V46: 35-42.
36. McGovern RH, Feddes JJ, Robinson FE, Hanson JA (1999) Growth performance, carcass characteristics, and the incidence of ascites in broilers in response to feed restriction and litter oiling. Poult Sci 78: 522-528.
37. Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R (2002) ASReml User Guide Release 1.0. 5 The Waterhouse, Waterhouse St, Hemel Hempstead, HP11ES, UK: VSN International,.
38. Mosaik Assembler. http://bioinformatics.bc.edu/marthlab/Mosaik.

Accessed Dec 2010.

39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
40. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research 38: e164.
41. Ensembl Genome Browser www.ensembl.org
42. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protocol 4: 1073-1081.
43. International Chicken Polymorphism Map Consortium (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432: 717 - 722.
44. Druyan S, Ben-David A, Cahaner A (2007) Development of Ascites-Resistant and Ascites-Susceptible Broiler Lines. Poult Sci 86: 811-822.
45. Druyan S, Cahaner A (2007) Segregation Among Test-Cross Progeny Suggests That Two Complementary Dominant Genes Explain the Difference Between Ascites-Resistant and Ascites-Susceptible Broiler Lines. Poult Sci 86: 2295-2300.
46. Navarro P, Visscher PM, Chatziplis D, Koerhuis ANM, Haley CS (2006) Segregation analysis of blood oxygen saturation in broilers suggests a major gene influence on ascites. British Poultry Science 47: 671 - 684.
47. Hillier L, Miller W, Birney E, Warren W, Hardison R, *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695 - 716.
48. Elferink M, van As P, Veenendaal T, Crooijmans R, Groenen M (2010) Regional differences in recombination hotspots between two chicken populations. BMC Genetics 11: 11.
49. Groenen M, Wahlberg P, Foglio M, Cheng H, Megens H, *et al.* (2009) A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res 19: 510 - 519.
50. Megens H-J, Crooijmans R, Bastiaansen J, Kerstens H, Coster A, *et al.* (2009) Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genetics 10: 86.
51. Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12: 32-42.

52. Van Laere A-S, Nguyen M, Braunschweig M, Nezer C, Collette C, *et al.* (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425: 832-836.
53. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470-476.
54. 1000 Genomes Project Consortium et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
55. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare Variants Create Synthetic Genome-Wide Associations. PLoS Biol 8: e1000294.
56. Dominguez MG, Hughes VC, Pan L, Simmons M, Daly C, *et al.* (2007) Vascular endothelial tyrosine phosphatase (VE-PTP)-null mice undergo vasculogenesis but die embryonically because of defects in angiogenesis. Proceedings of the National Academy of Sciences 104: 3243-3248.
57. Hassanzadeh M, Bozorgmehri Fard M, Buyse J, Bruggeman V, Decuypere E (2004) Effect of chronic hypoxia during embryonic development on physiological functioning and on hatching and post-hatching parameters related to ascites syndrome in broiler chickens. Avian Pathol Dec: 558-564.
58. Baumer S, Keller L, Holtmann A, Funke R, August B, *et al.* (2006) Vascular endothelial cell-specific phosphotyrosine phosphatase (VE-PTP) activity is required for blood vessel development. Blood 107: 4754-4762.
59. Mulvany MJ (1993) Vascular remodelling in hypertension. European Heart Journal 14: 2-4.
60. Intengan HD, Schiffrin EL (2001) Vascular Remodeling in Hypertension: Roles of Apoptosis, Inflammation, and Fibrosis. Hypertension 38: 581-587.
61. Holmes DIR, Wahab NA, Mason RM (1999) Cloning and Characterization of ZNF236, a Glucose-Regulated Kruppel-like Zinc-Finger Gene Mapping to Human Chromosome 18q22-q23. Genomics 60: 105-109.
62. Barrett J, Fry B, Maller J, Daly M (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263 - 265.
63. Wigginton JE, Cutler DJ, Abecasis GR (2005) A Note on Exact Tests of Hardy-Weinberg Equilibrium. The American Journal of Human Genetics 76: 887-893.
64. Hassel D, Dahme T, Erdmann J, Meder B, Huge A, *et al.* (2009) Nexilin mutations destabilize cardiac Z-disks and lead to dilated cardiomyopathy. Nat. Med. 15: 1281-1288.

65. Wang H, Li Z, Wang J, Sun K, Cui Q, *et al.* (2010) Mutations in NEXN, a Z-Disc Gene, Are Associated with Hypertrophic Cardiomyopathy. The American Journal of Human Genetics 87: 687-693.
66. Ramaraj R (2008) Hypertrophic Cardiomyopathy: Etiology, Diagnosis, and Treatment. Cardiology in Review 16: 172-180.
67. O'Keefe LV, Colella A, Dayan S, Chen Q, Choo A, *et al.* (2011) Drosophila orthologue of WWOX, the chromosomal fragile site FRA16D tumour suppressor gene, functions in aerobic metabolism and regulates reactive oxygen species. Human Molecular Genetics 20: 497-509.
68. Bottje WG, Wideman RF, Jr. (1995) Potential role of free radicals in the pathogenesis of pulmonary hypertension syndrome. Poultry and Avian Biology Reviews 6: 211-231.

5

Signatures of selection in the genome of commercial and non-commercial chicken breeds

M.G. Elferink[1], H-J. Megens[1], A. Vereijken[2], X. Hu[3], R.P.M.A. Crooijmans[1], M.A.M. Groenen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University and Research Centre, Marijkeweg 40, PO Box 338, Wageningen, The Netherlands;
[2]Hendrix Genetics Research, Technology & Services B.V., Spoorstraat 69, PO Box 114, Boxmeer, the Netherlands;
[3] State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing, China

Submitted

Abstract

The identification of genomic regions that have undergone selection will aid in understanding the domestication and selection history of the domesticated chicken. Furthermore, the identification of important genes underlying these regions of selection will aid in improvement of production traits and disease resistance. In the current study, we aimed to make a broad assessment of the effects of selection histories in domesticated chicken. Towards this end, we sampled commercial chickens representing all major breeding goals from multiple breeding companies. In addition, we sampled non-commercial chicken diversity by sampling almost all recognized traditional breeds The Netherlands, and a representative sample of breeds from China. This broad sample of 67 commercial and non-commercial breeds was assessed for signatures of selection in the genome using information of 57,636 SNPs that were genotyped on pooled DNA samples. We identified 396 regions of putative selection within the chicken genome of which 26 show strong evidence of selection. Our approach demonstrates the strength of including many different populations with similar, and breed groups with different selection histories to reduce stochastic effects based on single populations. The detection of the regions of putative selection resulted in the identification of several candidate genes that could aid in further improvement of production traits and disease resistance.

5.1 Introduction

The domesticated chicken exhibits a large variety of phenotypes differing in morphology, physiology and behavior [1]. Traditional breeds are nowadays mostly kept for ornamental purposes and show a large diversity in morphological phenotypes between breeds. Charles Darwin already noticed the large diversity of phenotypes within the chicken and assumed a single-origin for the domesticated chicken descending from *Gallus gallus* (Red Junglefowl) [2]. Although the single-origin was supported by many studies (e.g. [3,4,5,6]), it was debated by others [7,8]. Molecular genetic evidence supports multiple instances and multiple regions of domestication of the chicken from Red Junglefowl. Moreover, recent evidence supports genetic contributions from other Junglefowl species to current domesticated chickens. For instance, the yellow skin locus present in several domestic chicken breeds most likely originated from the *Gallus sonneratii* (Grey Junglefowl) [9]. Archeological findings, moreover, suggest that multiple domestication events were involved in the establishment of the domesticated chicken [10,11,12].

The chicken may initially not have been domesticated as a new food resource, but mainly for cultural reasons such as religion, decoration, and cock fighting [13]. Although selective breeding of chickens as a food resource has been documented to occur by the time of the Roman Empire [13], the strongest artificial selection most likely took place in the 20th century by commercial breeding companies. Specialized breeding lines, intensely selected on either growth traits (meat production) or reproductive traits (egg-laying) led to a massive selection response to those breeding goals [13,14,15]. The vast majority of chickens alive today in Europe and Northern America are bred for commercial purposes and are derived from only a handful of breeds. Although non-commercial breeds are still present, effective population sizes generally are estimated to be very small, and many breeds are threatened with inbreeding or extinction, thereby enhancing the loss of biodiversity in chicken [6].

Domestication of the chicken resulted in population bottlenecks, population growth, admixture of populations, inbreeding, genetic drift, and selective breeding. As a consequence of these events the genetic variation within the chicken genome must have changed from its ancestral state. Selection on desirable alleles will lead to a reduction or loss in nucleotide diversity at and near the selected locus, often referred to hitch-hiking or selective sweep [16,17]. The progress of production traits made in commercial breeds due to selective breeding has unfortunately also led to an increase in the occurrence of undesirable traits and diseases such as

reduced resistance to infectious disease [18], skeletal deformities [19], osteoporosis [20], and the pulmonary hypertension syndrome [21,22,23,24]. These undesirable traits and diseases may be the result of negative pleiotropic effects of the alleles under selection or from genetic hitch-hiking of undesirable alleles with the alleles under selection. To better understand these hitch-hiking effects on genetic diversity and negative pleiotropy, it is essential to identify regions and genes that have undergone a selective sweep. Furthermore, this information should aid in understanding the domestication and selection history of the domesticated chicken, and how molecular pathways may have been altered compared to the ancestral state, thereby facilitating the discovery of important genes and further improvement of production traits.

A recent study identified regions and genes putatively under selection during chicken domestication using a massive parallel sequencing strategy [1]. This study, however, only focused on a small number of breeds, making generalizations on selection history throughout the domesticated and wild chickens uncertain.

In the current study, we aimed to make a broad assessment of the effects of selection histories in domesticated chicken. Towards this end, we sampled commercial chickens representing all major breeding goals from multiple breeding companies. In addition, we sampled non-commercial chicken diversity by sampling almost all recognized traditional breeds from a Western-European country (The Netherlands), and a representative sample of breeds from China. In addition, several non-domesticated chicken populations were sampled, as well as related non-domesticated species (Gallus lafayetii). This broad sample of 67 commercial and non-commercial breeds was assessed for signatures of selection in the genome using information of 57,636 SNPs that were genotyped on pooled DNA samples. Having multiple populations for each breed should aid in decreasing the influence of stochastic effects such as genetic drift that may result from using just a single population. Furthermore, this strategy may reveal larger scale breed or breeding goal specific selection histories, rather than population-specific selection histories, potentially making it easier to interpret signatures of selection.

5.2 Materials and Methods

Study breeds

Individual blood samples were collected from 67 different chicken breeds varying from 8 to 75 individuals per breed (Table 5.1). Pools were made by either adding equal amounts of blood before DNA extraction, or by adding equal amounts of DNA after extraction, for each individual within each breed. DNA concentrations were measured by a NanoDrop spectrophotometer. Pooled DNA samples can be used to

calculate allele frequencies of SNPs [25,26]. Final DNA concentration of the pooled samples was 50-100 ng/ul.

The 67 breeds represent multiple populations of commercial broiler dam (n=5) and sire (n=8) lines, commercial white (n=11) and brown (n=11) egg-layers, Dutch traditional breeds (n=19), and Chinese breeds (n=10) (Table 5.1). Two subspecies from *Gallus gallus* (*Gallus gallus gallus, Gallus gallus spadiceus*) were also included as well as the *Gallus lafayetii* that was used as an outgroup (Table 5.1).

Table 5.1 Information on the breeds genotyped.

	Breed name	# ind[1]	Hp[2]	Origin[3]	Breed groups[4]			
Junglefowls	*G. lafayetii*	11	0.04	Sri Lanka	Outgroup			
	G. g. gallus[5]	30	0.37	Thailand	NDM			
	G. g. spadiceus[5]	30	0.36	Thailand	NDM			
Broiler sire line	Broiler sire 1	75	0.42	commercial	DM	CM	BR	BRS
	Broiler sire 2	75	0.43	commercial	DM	CM	BR	BRS
	Broiler sire 3	75	0.43	commercial	DM	CM	BR	BRS
	Broiler sire 4	75	0.42	commercial	DM	CM	BR	BRS
	Broiler sire 5	75	0.39	commercial	DM	CM	BR	BRS
	Broiler sire 6	75	0.41	commercial	DM	CM	BR	BRS
	Broiler sire 7	75	0.42	commercial	DM	CM	BR	BRS
	Broiler sire 8	48	0.39	commercial	DM	CM	BR	BRS
Broiler dam line	Broiler dam 1	75	0.36	commercial	DM	CM	BR	BRD
	Broiler dam 2	75	0.35	commercial	DM	CM	BR	BRD
	Broiler dam 3	75	0.40	commercial	DM	CM	BR	BRD
	Broiler dam 4	75	0.41	commercial	DM	CM	BR	BRD
	Broiler dam 5	75	0.42	commercial	DM	CM	BR	BRD
White egg-layer	White layer 1	75	0.24	commercial	DM	CM	LR	WL
	White layer 2	75	0.27	commercial	DM	CM	LR	WL
	White layer 3	75	0.26	commercial	DM	CM	LR	WL
	White layer 4	75	0.25	commercial	DM	CM	LR	WL
	White layer 5	75	0.28	commercial	DM	CM	LR	WL
	White layer 6	75	0.21	commercial	DM	CM	LR	WL
	White layer 7	75	0.26	commercial	DM	CM	LR	WL
	White layer 8	75	0.29	commercial	DM	CM	LR	WL
	White layer 9	75	0.27	commercial	DM	CM	LR	WL
	White layer 10	75	0.22	commercial	DM	CM	LR	WL
	White layer 11	75	0.28	commercial	DM	CM	LR	WL
Brown egg-layer	Brown layer 1	75	0.31	commercial	DM	CM	LR	BL
	Brown layer 2	75	0.32	commercial	DM	CM	LR	BL
	Brown layer 3	75	0.32	commercial	DM	CM	LR	BL
	Brown layer 4	75	0.31	commercial	DM	CM	LR	BL
	Brown layer 5	75	0.37	commercial	DM	CM	LR	BL
	Brown layer 6	75	0.31	commercial	DM	CM	LR	BL
	Brown layer 7	75	0.32	commercial	DM	CM	LR	BL
	Brown layer 8	75	0.35	commercial	DM	CM	LR	BL
	Brown layer 9	75	0.32	commercial	DM	CM	LR	BL
	Brown layer 10	75	0.34	commercial	DM	CM	LR	BL
	Brown layer 11	75	0.32	commercial	DM	CM	LR	BL

Table 5.1 Continued…

	Breed name	# ind[1]	Hp[2]	Origin[3]	Breed groups[4]			
Dutch	Groninger mew bantam	21	0.30	the Netherlands	DM	NCM	DU	DCF
	Groninger mew	22	0.28	the Netherlands	DM	NCM	DU	DCF
	Lakenvelder	46	0.27	the Netherlands	DM	NCM	DU	DCF
	Drente fowl	13	0.33	the Netherlands	DM	NCM	DU	DCF
	Assendelf fowl	22	0.28	the Netherlands	DM	NCM	DU	DCF
	Friesian fowl	9	0.33	the Netherlands	DM	NCM	DU	DCF
	Hamburgh	50	0.30	the Netherlands	DM	NCM	DU	DCF
	Polish bearded	30	0.24	the Netherlands	DM	NCM	DU	DPB
	Owl-bearded Dutch	8	0.33	the Netherlands	DM	NCM	DU	DPB
	Polish non-bearded	49	0.16	the Netherlands	DM	NCM	DU	DPB
	Breda fowl	13	0.33	the Netherlands	DM	NCM	DU	DPB
	Brabanter	50	0.34	the Netherlands	DM	NCM	DU	DPB
	Dutch bantam	23	0.32	the Netherlands	DM	NCM	DU	DPB
	Booted bantam	12	0.32	the Netherlands	DM	NCM	DU	DPB
	Barnevelder	11	0.29	the Netherlands	DM	NCM	DU	DNB
	Welsumer	41	0.31	the Netherlands	DM	NCM	DU	DNB
	North-Holland blue	34	0.33	the Netherlands	DM	NCM	DU	DNB
	Kraienkoppe	48	0.32	the Netherlands	DM	NCM	DU	DNB
	Schijndelaar	12	0.33	the Netherlands	DM	NCM	DU	DNB
Chinese	Bian	21	0.41	China (In. Mongolia)	DM	NCM	CH	
	Chahua	34	0.33	China (Yunnan)	DM	NCM	CH	
	Chongren Ma	40	0.35	China (Jiangxi)	DM	NCM	CH	
	Henan Game	25	0.33	China (Henan)	DM	NCM	CH	
	Gushi	29	0.36	China (Henan)	DM	NCM	CH	
	Luyuan	30	0.38	China (Jiangsu)	DM	NCM	CH	
	Wenchang	35	0.42	China (Hainan)	DM	NCM	CH	
	Wahui	32	0.41	China (Jiangxi)	DM	NCM	CH	
	Xianju	48	0.36	China (Zhejiang)	DM	NCM	CH	
	Xiaoshan	32	0.38	China (Zhejiang)	DM	NCM	CH	

[1] Number of individuals in genotyped DNA pool. [2] Average Hp based on all markers. [3] Name of country (province) of origin. [4] Breed groups for the breeds, DM =domesticated, NDM = non-domesticated, CM= commercial, NCM= non-commercial, BR= broiler, LR= layer, DU= Dutch, CH= Chinese, BRS= broiler sire line, BRD= broiler dam line, WL= white egg-layer, BL= brown egg-layer, DCF= Dutch countryfowls, DPB= Dutch polish and bearded, and DNB= Dutch new breeds. [5] These breeds are part of the AvianDiv project [6]. *G. g. gallus* = Aviandiv102 and *G. g. spadiceus* = Aviandiv101.

Marker selection and allele frequency calculations

In total, 57,636 SNPs were included on the Illumina Infinium iSelect Beadchip (Table S1). For GGA1–GGA5 and GGAZ, markers were selected every 15 kb; for GGA6–GGA10 every 10 kb; for GGA11–GGA20 every 7.5 kb; and for GGA21–GGA28, GGAW and the two linkage groups LGE22C19W28_E50C23 (reffered to as LGE22) and LGE64, every 5 kb. Genotyping was performed using the standard protocol for Infinium iSelect Beadchips and raw data were analyzed with GenomeStudio v2009.2. Markers with a normalized R value of less than 0.15 were not included in

further analysis. For the DNA pools, the normalized allele frequency $\hat{p}_n$ was calculated by combining the heterozygote correction equation of Hoogendoorn *et al.* [27] with the "normalization 4" equation of Peiris *et al.* [25]; $\hat{p}_n = \frac{\left(\frac{X_{RAW}}{X_{RAW} + \kappa Y_{RAW}}\right) - \hat{\beta}_0}{\hat{\beta}_1}$, where X_{RAW} is the raw intensity of allele A, X_{RAW} is the raw intensity of allele B, and κ is the ratio of the average X_{RAW} and Y_{RAW} intensities based on heterozygote individuals. $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1$ is the slope of a simple linear regression of the observed mean heterozygote-corrected frequencies based on individuals with genotype AA, AB and BB on their expected frequencies of 1, 0.5 and 0, respectively. κ, $\hat{\beta}_0$ and $\hat{\beta}_1$ values were calculated from a panel of 458 individuals, including white and brown egg-layers, broilers, Dutch traditional breeds, *Gallus gallus spadiceus*, and *Gallus lafayetii*. If the heterozygous genotype class was missing, heterozygote correction was not performed and κ was set to 1. SNPs that were homozygous in all individual animals were removed from the data. To avoid genotype mistakes made due to technical errors, a genotype class had to contain at least three individual animals to be included in the calculation of κ, $\hat{\beta}_0$ and $\hat{\beta}_1$. Animals from the *Gallus lafayetii* were genotyped individually and genotypes were pooled in silico to estimate allele frequencies for this population.

Genetic distance calculations

PHYLIP software (version 3.69; [28]) was used to calculate pairwise genetic distances between the breeds. Nei genetic distance was used as a measure for genetic distance [29]. Because PHYLIP is unable to deal with missing data, distance calculations for each pair of breeds were based on the marker data that these breeds had in common [30]. Mega 4.0 software [31] was used for hierarchical clustering using the Neighbor-Joining procedures on the genetic distance matrix for all breeds. *Gallus lafayetii* was used to root the tree.

Signatures of selection

To decrease the influence of stochastic effects such as genetic drift, analysis on signatures of selection were performed on pooled data of groups of breeds. The breeds were grouped in fourteen different breed groups in four levels (Table 5.1). The first level included all domesticated breeds (DM, n=64). The two non-

domesticated breeds were not grouped and analyzed because the group size was too small. The second level was based on their commercial background and included commercial (CM, n=35) and non-commercial (NCM, n=29) breeds. The third level was based on either their general commercial purpose or geographical location and included broiler (BR, n=13), layer (LR, n=22), Dutch traditional (DU, n=19) and Chinese (CH, n=10) breeds. The fourth level was based on either their position in the dendrogram (Figure 5.1) and included the broiler sire lines (BRS, n=8), broiler dam lines (BRD, n=5), white egg-layers (WL, n=11) and brown egg-layers (BL, n=11), or were based on their classical classification and included the Dutch countryfowls (DCF, n= 8), Dutch polish and bearded (DPB, n= 5), and Dutch new breeds (DNB, n=6).

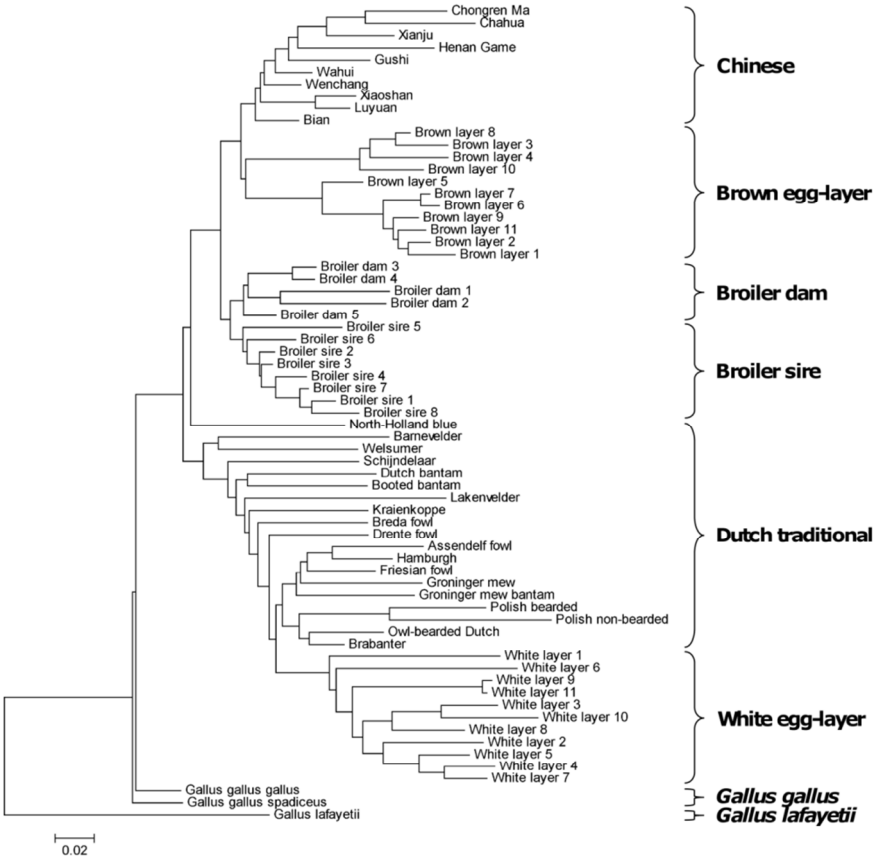


Figure 5.1 Dendrogram for the 67 breeds based on Nei genetic distance. Accolades indicate the breed groups for the clusters as used in this study.

To identify regions under selection the "Z transformed heterozygosity" (ZHp) approach of Rubin *et al.* [1] was used. In an overlapping sliding window approach the heterozygosity H_p was calculated as: $H_p = \frac{2\sum n_{MAJ} \sum n_{MIN}}{\left(\sum n_{MAJ} + \sum n_{MIN}\right)^2}$, where $\sum n_{MAJ}$ is the sum of major allele frequencies, and $\sum n_{MIN}$ is the sum of the minor allele frequency within a window. Individual H_p values were Z-transformed: $ZH_p = \frac{(H_p - \mu H_p)}{\sigma H_p}$, where μH_p is the overall average heterozygosity and σH_p is the standard deviation for all windows within one breed. In this study we focus on the extreme lower end of the ZHp distribution and considered windows with ZHp values less or equal to -6 to show strong evidence of selection. Windows with ZHp values less or equal to -4 were considered to show weak evidence of selection.

5.3 Results

From the 57,636 SNPs that were included on the chip, 51,080 were used for analysis. Because the breed pools included both female and male individuals, the analysis were only performed on autosomal markers and therefore the 3,023 markers located on chromosome W and Z were excluded from the analysis. Moreover, 1,144 markers were unmapped on the current genome build and were therefore also excluded from the analysis. A total of 2,389 markers were excluded as they were either homozygous in all individual animals (n= 2,146) or did not pass the quality control (n= 243).

The 51,080 autosomal SNPs were used to construct a tree representing genetic distances between 67 breeds (Figure 5.1). The two RJF subspecies and *Gallus lafayetii* cluster separate from the domesticated breeds. The domesticated breeds are divided in two branches. Brown egg-layers, broilers and Chinese breeds cluster together in one branch, while white-egg layers and Dutch traditional breeds cluster in the other. Within the broiler cluster, a clear distinction was found between the broiler sire and broiler dam lines. The Dutch traditional breeds cluster together according to their classical classification with a few exceptions.

To identify regions that are likely to be or have been under selection, Hp and ZHp values were calculated for a number of different marker window sizes for all fourteen breed groups (data not shown). Based on these analyses we decided to

focus on a size of 5 markers per window (Figure 5.2, Figure 5.3, Table S2). This size enabled us to obtain a normal distribution for the Hp values (Figure S1) while still remaining a relatively high resolution to detect candidate genes. Moreover, a marker window size of five enabled us to detect the empirically proven selective sweep at the *BCDO2* locus [9] while an increased number of markers per window resulted in the loss of detection of this locus. The ZHp threshold values for weak (ZHp less or equal to -4) and strong (ZHp less or equal to -6) evidence were chosen because these represent the extreme lower end of the distribution (Figure S2).
After merging consecutive windows, 396 regions were identified where at least one breed group showed weak evidence of selection (Table S3). In total, 26 regions showed strong evidence of selection (Table 5.2). Three of these regions (R11, R25, and R26) were found exclusively within in the broiler breed groups. All three showed strong evidence of selection in the broiler sire line and R11 also showed weak evidence in the broiler dam line. Region R1 showed strong evidence for selection exclusively within the broiler sire breed group. Region R8 showed strong evidence of selection exclusively within the Chinese breed group. Linkage group LGE64 consisted of only 4 markers and was not included in further analysis. The average overall heterozygosity (μH_p) and standard deviation (σH_p) for the fourteen breed groups are shown in Supplementary Table 4 (Table S4). The average heterozygosity for each breed based on all markers is shown in Table 5.1. Average sizes for the 5 markers windows were; 97 kb for GGA1-5, 71 kb for GGA6-10, 46 kb for GGA11-20, and 31 kb for GGA21-GGA28 and linkage groups LGE22.
With a marker window size of five, no region with strong evidence of selection was identified in the Dutch (DU, DCF, DPB, and DNB), white egg-layers (WL) and brown egg-layer (BL) breed groups. Increasing the number of markers per window resulted in the detection of regions with strong evidence in these breed groups although these regions were large and contained many genes (data not shown). For example, we detected one region with strong evidence of selection in the white egg-layer breed group at a window size of 110 markers. This region on chromosome 20 (5.32-7.69 Mb) is 2.36 Mb in size and contains 46 genes. Note that this region is also observed in the five marker window data as a continuous stretch of low ZHp values (Figure 5.3).

5.4 Discussion

The position of the breeds in the dendrogram (Figure 5.1) is largely in agreement with previous published data [32]. Our data, moreover, fits with the putative historical origin of the breeds. The broilers, and brown egg-layers cluster between

the Chinese breeds on one side and the Dutch and white egg-layers on the other side. The broiler and brown egg-layer breeds were established in the late 19th and early 20th century by crossing European breeds with Asian breeds [13] [33] and molecular evidence of this Asian introgression was published recently [34]. The Dutch traditional and white egg-layer breeds both have their origin in Europe [13,33]. We used an in silico pooling approach of populations, defining groups based on overall genetic relatedness to decrease stochastic effects such as genetic drift in our analysis. If a region under selection is present in only one breed, it will be averaged out due to a high diversity in the other breeds included in the same breed group. However, if a region is present in all breeds, the confidence that this region is truly under selection will increase.

Although we identified regions of strong selection within most breed groups we were not able to identify these regions within the Dutch breed groups (either separate or in the classification breed groups) nor did we in the white- and brown egg-layer breed groups. For the Dutch breeds, we might have been unable to detect regions with strong evidence of selection as these breeds are genetically too diverse. Each Dutch breed has been intensely selected for specific phenotypic characteristics and there might be very little overlap in selected regions between the breeds within the breed group. The lack of identification of regions under selection within the white and brown egg-layers breed groups most likely results from the origin of the breeds. Both the white and brown egg-layers were created using a small base population and this founder effect resulted in a major population bottleneck [33]. Regions with low genetic diversity caused by the bottleneck will exist in all breeds derived from the base population. Our method determines if the heterozygosity of a given marker window is an outlier compared to the average heterozygosity of the genome. Because the existence of many low diversity regions will lower the average heterozygosity and increase the standard deviation, we were not able to detect outlier genomic regions. To decrease the standard deviation we analyzed our data with an increased number of markers per windows. Although this indeed resulted in the detection of regions with strong evidence of selection within the Dutch, brown and white egg-layer breed groups, these regions were generally large and contained many genes. Because of the decreased resolution to detect candidate genes we decided to focus on small window size only. Five markers per window provided a good resolution to detect genes while retaining a normal distribution for the Hp values (Figure S1).

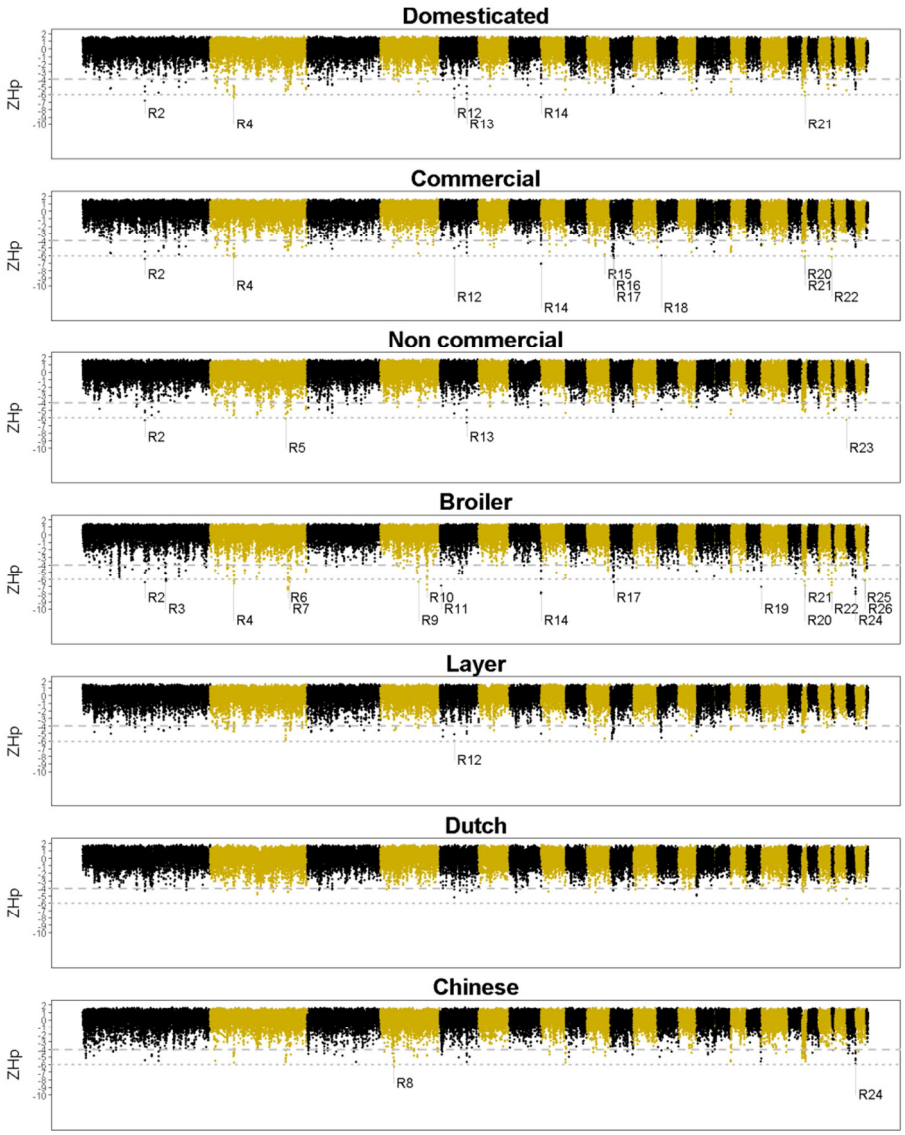


Figure 5.2 ZHp values for seven breed groups using a sliding window of five markers across the genome. Odd chromosomes numbers (and LGE22) are shown in black and even chromosome numbers are shown in yellow. The grey dotted line indicates a ZHp threshold value of -4 or -6. For the regions with strong evidence of selection the ID is shown beneath the plot.

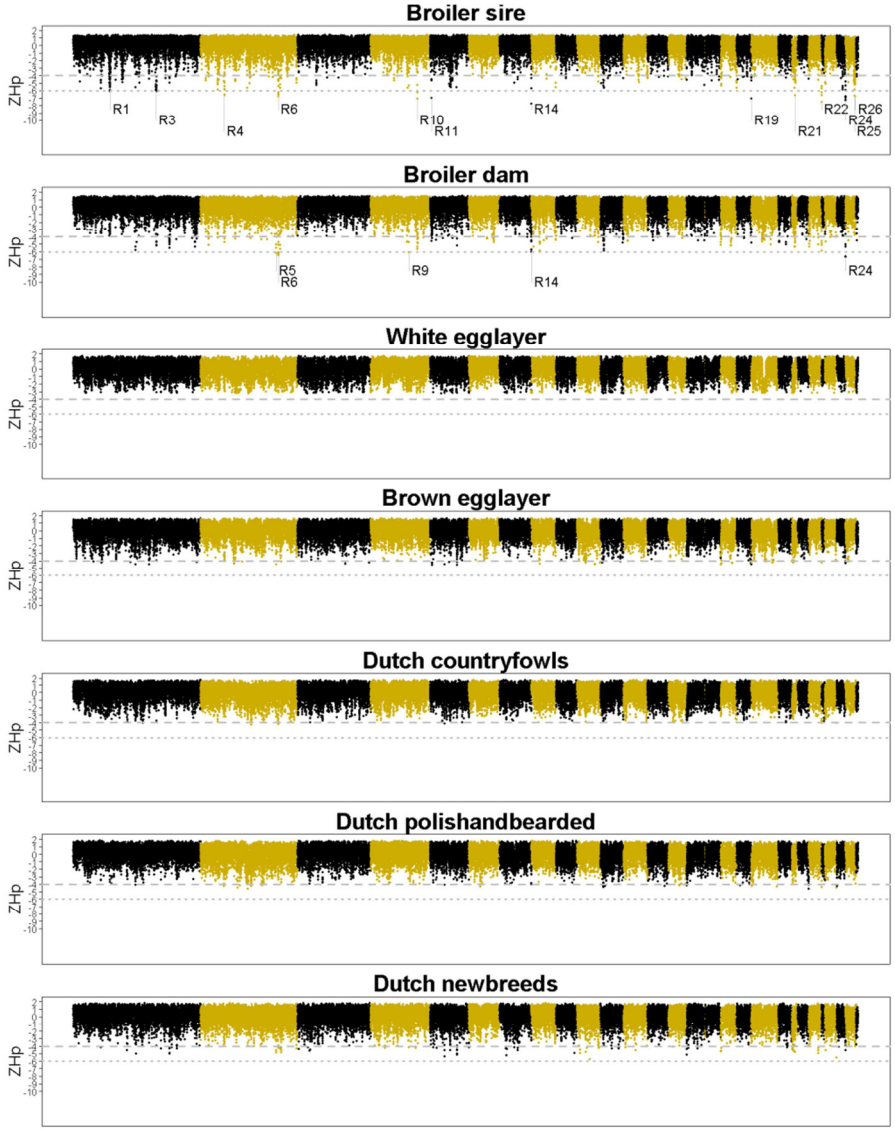


Figure 5.3 ZHp values for seven breed groups using a sliding window of five markers across the genome. Odd chromosomes numbers (and LGE22) are shown in black and even chromosome numbers are shown in yellow. The grey dotted line indicates a ZHp threshold value of -4 or -6. For the regions with strong evidence of selection the ID is shown beneath the plot.

In order to detect regions under selection in the egg-layers, we combined the two breed groups of the white and brown egg-layers. Both egg-layers have been selected for similar production traits related to egg production and therefore combining these two breeds groups could lead to the identification of the same genomic regions independently being selected for similar egg production traits.
Of the 26 regions that show strong evidence of selection (Table S3), 13 were previously described [1]. The identification of these regions in two independent studies with various detection methods implies that these regions have strong signatures of selection and are likely to be true positives. Some of these regions contain genes with biological functions that were previously related to traits under selection in the chicken. Region R1 includes insulin-like growth factor 1 (*IGF1*) that is associated with growth, body composition, and skeleton integrity in chicken [35,36], and pro-melanin-concentrating hormone (*PMCH*) that is involved in body weight and feed intake in mice [37]. The gene encoding beta-carotene oxygenase 2 (*BCDO2*), which is involved in the yellow skin phenotype in chicken [9], is found in R22 and has strong evidence of selection in commercial breeds. In addition, several other genes with biological functions that could be related to (production) traits are found in other regions of putative selection. Region R6, detected in broiler breed groups, contains the *HNF4G* gene. *HNF4G* -/- knockout mice have a significant higher bodyweight at 7 weeks compared to normal mice, presumably caused by decreased energy expenditure that results from a reduced locomotor activity [38]. Moreover, feed and water intake is also significantly lower in the knockout mice [38]. The biological function suggests that selection on *HNF4G* might be involved in the selection on bodyweight or feed conversion ratio in the chicken. This view is strengthened by the fact that the strongest selection was observed in broilers (sire as well as dam lines) (Table 5.2). Moreover, a significant QTL for bodyweight overlaps with region R6 and was found to explain about 7% of bodyweight in broilers from 3 to 7 weeks of age [39]. Region R11, detected in broiler sire (strong evidence) and broiler dam (weak evidence), is embedded within the gene encoding NEL-like 1 (*NELL1*). *NELL1* is involved in bone tissue formation and *NELL1*-deficient mice have skeletal defects in the cranial vault, vertebral column and ribcage [40,41]. The biological functions of *NELL1*, combined with the evidence of selection, might relate to selection on the skeletal integrity of modern broilers. Skeletal integrity might have been co-selected with growth rate and meat yield as the skeleton had to support the weight of the modern broiler [42]. Animals not capable of dealing with the increasing bodyweight could develop defects such as tibial dyschondroplasia, valgus-varus deformity, and spondylolisthesis [19] and will be rejected from the breeding program. This rejection will essentially lead to a

positive selection for skeletal integrity. Heavy birds are more prone to develop these leg-problems and therefore it is expected that selection will have been strongest in the heavier breeds. This is agreement with region R11 shows strong evidence of selection in the heavy broiler sire lines and weak evidence in the slightly less heavy broiler dam lines. In previous studies no significant QTLs associated with bone or skeletal traits were found near region R11 [43].

Besides the 13 regions that were previously described we also identified 13 additional regions with strong evidence of selection. On the other hand, several regions with strong evidence of selection found previously [1] are not identified in our study. These differences in the identified regions may in part result from different methodological approaches used in both studies. While our study was based on many breeds genotyped with a SNP genotyping assay, the study of Rubin *et al.* [1] was based on low coverage whole genome re-sequencing of a small number of breeds. Regions detected in our study might be poorly covered in the massive parallel sequencing strategy or might have been not be detected simply because the breeds were not included. In addition, we included more breeds per breed group that might result in less false positive regions found as a result of genetic drift.

While the approach described in this study has several strong advantages – the ability to include many different populations cost-effectively being among the most important – the application of SNP based assays has limitations, notably ascertainment bias and low marker resolution. The SNPs selected in our study were discovered in two independent studies. One study is based on comparing the *G. gallus* genome sequence derived from a single RJF to that of one Silkie, one white egg-layer, or one broiler [44]. The second study is based on massively parallel sequencing of four pools of commercial chicken (two broiler lines, a white egg layer line, and a brown egg layer line (MAMG, unpublished). A SNP was discovered when a single nucleotide polymorphism was observed between the reference RJF and one of the four discovery breeds. Therefore, it is possible that breed, or animal, specific SNPs were selected for the genotyping assay. Breed specific markers will not segregate in other breeds, thereby resulting in a false positive signal of selection in the other breeds. The selection of markers that eventually are included in the genotyping assay also introduces a bias. Selection criteria for the SNPs are mainly based on their minor allele frequency (MAF) and position on the genome. SNPs that are near fixation in the four SNP discovery breeds will have a low MAF, also when they are nearly fixed for the non-reference allele. Because all four SNP discovery breeds represent domesticated breeds, particularly regions under

selection due to domestication will be underrepresented since SNPs within these regions will have low MAF and are not included in the genotyping assay.

Table 5.2 Regions of putative selection identified in this study.

ID	Chr	Position (Mb)[1]	DM	CM	NCM	BR	LR	DU	CH	BRS	BRD	WL	BL	DCF	DPB	DNB
R1	1	57.16-57.64	-3.5	-3.2	-3.3	**-5.8**	-1.2	-3.8	-2.7	**-6.1**	-3.7	-0.9	-2.2	-2.6	-3.7	-3.3
R2	1	98.80-98.95	**-6.8**	**-6.5**	**-6.3**	**-6.5**	-4.7	-4.4	-5.4	-5.6	-5.8	-1.7	-4.5	-3.7	-4.1	-3.6
R3	1	131.15-131.59	-3.6	-4.1	-2.9	**-6.3**	-2.1	-3.5	-2.9	**-6.1**	-5.2	-3.1	-2.1	-3.7	-2.0	-4.0
R4	2	35.22-35.69	**-6.5**	**-6.3**	-5.7	**-6.6**	-4.7	-3.7	-5.8	**-6.6**	-4.8	-2.8	-3.9	-3.2	-4.3	-3.8
R5	2	120.56-120.87	-5.6	-4.9	**-6.1**	-2.0	-5.9	-4.0	-5.7	-0.7	**-6.2**	-2.9	-4.5	-3.2	-2.3	-4.9
R6	2	123.46-123.90	-5.3	-5.3	-4.6	**-7.4**	-3.1	-4.6	-2.2	**-6.7**	**-6.5**	-3.1	-1.1	-4.2	-3.4	-4.3
R7	2	126.69-126.91	-4.6	-4.2	-4.4	**-6.4**	-2.1	-3.0	-3.7	-5.6	-5.6	-0.6	-3.3	-1.9	-2.2	-4.3
R8	4	22.27-22.47	-2.0	-1.4	-2.5	-2.3	-0.5	-1.1	**-6.3**	-2.0	-1.7	0.0	-1.8	-0.5	-1.9	0.0
R9	4	60.47-60.60	-5.7	-5.7	-4.6	**-6.4**	-3.9	-2.9	-4.0	-5.3	**-6.1**	-1.5	-3.2	-2.9	-1.7	-4.1
R10	4	73.63-73.88	-0.8	0.0	-1.6	**-7.4**	0.7	-0.8	-4.9	**-7.1**	-5.6	-0.5	0.0	-1.6	-1.0	-1.3
R11	5	2.34-2.51	-0.5	-2.0	1.0	**-6.9**	0.2	1.3	-2.9	**-7.0**	-4.6	0.4	0.9	1.2	1.3	0.1
R12	5	24.50-24.71	**-6.5**	**-6.1**	-5.4	-3.7	**-6.0**	-5.2	-2.2	-3.1	-3.4	-3.2	-4.5	-4.2	-4.1	-5.4
R13	5	44.05-44.33	**-6.6**	-5.7	**-6.7**	-3.9	-5.0	-4.5	-5.7	-2.2	-5.2	-1.9	-4.6	-3.4	-3.5	-5.2
R14	7	38.03-38.35	**-6.4**	**-7.2**	-4.4	**-7.9**	-5.0	-3.8	-2.9	**-7.8**	**-6.0**	-3.1	-3.4	-2.5	-4.0	-3.3
R15	10	17.52-17.59	-4.5	**-6.1**	-3.0	-4.0	-5.7	-2.6	-2.1	-3.1	-3.8	-2.9	-4.5	-1.7	-2.3	-2.4
R16	11	2.53-2.67	-5.7	**-6.1**	-4.3	-5.8	-4.7	-3.0	-3.4	-5.3	-4.6	-2.1	-3.9	-2.6	-2.0	-3.6
R17	11	3.37-3.55	-5.6	**-6.4**	-3.7	**-6.5**	-4.7	-3.1	-1.6	-5.6	-5.9	-2.4	-3.5	-2.6	-1.8	-3.9
R18	13	3.82-3.93	-5.9	**-6.0**	-4.9	-4.0	-5.6	-3.2	-4.2	-3.5	-3.4	-3.1	-3.8	-2.5	-2.1	-3.5
R19	19	9.81-9.92	-2.8	-4.5	-1.1	**-7.0**	-2.2	0.6	-5.3	**-7.1**	-4.8	0.6	-4.5	0.4	0.1	-1.7
R20	22	1.95-2.01	-5.7	**-6.3**	-3.8	**-6.3**	-4.7	-1.7	-5.0	-5.6	-5.4	-2.4	-3.5	-1.0	-1.0	-2.4
R21	22	2.15-2.22	**-6.2**	**-6.1**	-5.0	**-6.9**	-4.2	-3.6	-4.1	**-6.6**	-5.3	-2.6	-2.3	-2.6	-2.5	-4.8
R22	24	6.25-6.31	-1.6	**-6.2**	0.3	**-7.8**	-3.9	0.5	-2.0	**-7.5**	-5.9	-0.9	-4.2	-1.1	-2.2	-2.6
R23	26	5.01-5.09	-5.5	-4.2	**-6.3**	-3.8	-3.1	-5.4	-3.2	-2.9	-3.9	-0.9	-2.7	-4.1	-4.5	-5.6
R24	27	4.61-4.84	-5.3	-4.9	-5.0	**-8.0**	-3.2	-3.1	**-6.0**	**-7.3**	**-6.7**	-2.3	-4.3	-2.3	-2.0	-4.5
R25	28	3.75-3.80	-1.0	-1.3	-1.8	**-6.2**	-0.2	-1.5	-1.6	**-6.7**	-3.8	-1.4	-0.1	-1.5	-1.4	-1.2
R26	28	4.04-4.07	-2.6	-2.2	-2.6	**-6.2**	0.0	-1.3	-3.4	**-6.7**	-3.6	1.1	-1.0	-0.8	-0.2	-3.6

[1] Position based on chicken genome build WASHUC2. ZHp values are shown for each region for each breed group. Values in bold values are less than or equal to -6. DM =domesticated, CM= commercial, NCM= non-commercial, BR= broiler, LR= layer, DU= Dutch, CH= Chinese, BRS= broiler sire line, BRD= broiler dam line, WL= white egg-layer, BL= brown egg-layer, DCF= Dutch countryfowls, DPB= Dutch polish and bearded, and DNB= Dutch new breeds.

Although markers are selected evenly throughout the whole genome, the resolution of the assay will not be sufficient to identify all regions of selection. The genomic size of selective sweeps is positively correlated to selection pressure, and negatively with recombination rate. Genomic regions under strong and recent directional selection located in relatively lowly recombining regions of the genome (e.g. the macro-chromosomes in birds compared to the micro-chromosomes [45,46,47]) will be much more readily detectable. Although the *TSHR* selective sweep is fixed for almost 40 kb in most domestic breeds [1], we were unable to identify this locus. In our analysis this 40 kb region is covered by only a single SNP and although this SNP is fixed in almost all domesticated breeds, the window that included this SNP never reached significance as the other markers in the window are segregating in relatively high frequencies. Although the massive parallel strategy does not suffer from the ascertainment bias described above, the high costs of this method restrict the number of breeds that can be included in the analysis. In this study, we specifically choose the less expensive SNPs assays in order to increase the total number of breeds. Not only are we able to comment on a wide variety of breeds, the increased amount of breed within a breed group enabled us to decrease the influence of stochastic effects such as genetic drift.

In our data we identified five regions (R1, R8, R11, R25, and R26) that are specific for one breed group (Table 5.2). Because these regions are not subjected to the possible bias of breed specific markers (if that was the case, we would expect to see the signature of selection in all but one breed group) we consider these to be reliable. Two regions (R1 and R11) have already been discussed above. R8 shows strong evidence of selection and is specific for the Chinese breeds. R8 includes two genes; platelet derived growth factor C (*PDGFC*) and the glycine receptor beta subunit (*GLRB*). Platelet derived growth factors are major mitogens and stimulants of motility in mesenchymal cells [48,49]. Mesenchymal cells can differentiate into a variety of cell types including bone and fat cells. In mice, *PDGFC* is widely expressed in mesenchymal precursors and the myoblast of the smooth and skeletal muscles [50]. Knockout studies in mice demonstrate that *PDGFC* is essential for palatogenesis, a process that forms the palate (roof of the mouth) and separates the oral cavity from the nasal cavity [51]. *GLRB* is involved in an important fertilization event, the sperm acrosome reaction which is the process facilitating entry of the spermatozoa into the oocyte [52,53]. *GLRB* is also associated with the neurological disorder hyperekplexia (startle syndrome) in human [54] and myoclonus (involuntary twitching of muscle) in mice [55]. A number of QTL regions are found in the chicken QTL database that cover region R8, although none were identified specifically in Chinese breeds [43]. These QTL regions are associated with

Marek's disease [56], residual feed intake [57], fear-related behavior [58], feed conversion ratio [59], residual feed intake [59], creatine kinase level [60], shank length [61], tibia strength [61], conformation score [62], and thigh muscle weight [63] [64]. The QTL for shank length and tibia strength appear most interesting when compared to the biological function of *PDGFC*. R25 and R26 show strong evidence of selection in broilers, and specifically in broiler sire lines. R25 includes three genes: SIN3 homolog B (*SIN3B*), HAUS augmin-like complex, subunit 8 (*HAUS8*), and C3 and PZP-like alpha-2-macroglobulin domain containing 8 (*CPAMD8*). *SIN3B* is a global regulator of transcription [65] and is essential for embryonic development [66]. mSin3B-/- knockout studies in mice indicated that knockout embryos displayed growth retardation [67]. *HAUS8* is a microtubule-associated protein required for maintenance of spindle integrity and chromosomal stability in human cells [68]. *CPAMD8* is a member of the complement 3/alpha2-macroglobulin family of proteins that are involved in innate immunity and damage control [69]. Several QTL regions are found in the chicken QTL database [43] that cover region R25. These QTL regions are associated with abdominal fat weight [70], skin fat weight [70], tibia cortex width [61], body weight (21 and 42 days) [71], breast muscle weight [71], carcass weight [64], and right ventricular weight as percentage of body weight [24] Due to the broad biological functions of the genes located within R25, it is difficult to comment on a possible relation between the associated QTLs and the genes located within R25. Region R26 includes evolutionary conserved regions with unknown function but does not contain any known genes. Region S26 is closely linked to R25, and the QTLs described for R25 also cover region R26.

In conclusion, we identified 396 regions of putative selection within the chicken genome and 26 of these regions show strong evidence of selection in at least one of the fourteen breed groups. Our approach demonstrates the strength of including many different populations with similar, and breed groups with different selection histories to reduce stochastic effects based on single populations. The detection of the regions of putative selection resulted in the identification of several candidate genes that could aid in further improvement of production traits and disease resistance.

Acknowledgments

The authors would like to thank Bert Dibbits for performing the genotyping experiments and Pieter van As for his contribution on the data analysis and material collection. This study was part of “The characterisation of genes involved in

pulmonary hypertension syndrome in chicken" project funded by the Dutch Technology Foundation (STW). Project number 07106.

Author Contributions

Conceived and designed the study: MGE, HJM, RPMAC, and MAMG. Analyzed the data: MGE. Contributed to materials: AV and XH. Wrote the paper: MGE, HJM, RPMAC, and MAMG.

Supplementary files

Supplementary data files are available on:
http://library.wur.nl/WebQuery/edepot/163759

Figure S1. Distribution of Hp values for all windows of five markers.

Figure S2. Distribution of ZHp values for all windows of five markers.

Table S1. Information on all SNP markers used. Chromosomal locations are based on the position in the WASHUC2 build.

Table S2. Hp and ZHp values for all windows of five markers. Chromosomal locations are based on the position in the WASHUC2 build. DM =domesticated, CM= commercial, NCM= non-commercial, BR= broiler, LR= layer, DU= Dutch, CH= Chinese, BRS= broiler sire line, BRD= broiler dam line, WL= white egg-layer, BL= brown egg-layer, DCF= Dutch countryfowls, DPB= Dutch polish and bearded, and DNB= Dutch new breeds.

Table S3. All regions of putative selection found and their underlying genes. Chromosomal locations are based on the position in the WASHUC2 build. Size refers to the total size of the merged windows. # windows refer to the number of merged windows. Region ID refers to the region with strong evidence of selection as described in this manuscript. DM =domesticated, CM= commercial, NCM= non-commercial, BR= broiler, LR= layer, DU= Dutch, CH= Chinese, BRS= broiler sire line, BRD= broiler dam line, WL= white egg-layer, BL= brown egg-layer, DCF= Dutch countryfowls, DPB= Dutch polish and bearded, and DNB= Dutch new breeds. Values of Rubin *et al.* refers to ZHP values found in a previous study [1]. The 'genes' column include information of the genes included in the region of putative

selection. For each gene the location within the region is given followed by the Ensembl chicken ID and human orthologs name if known. (1) gene is located within region, (2) region is located within gene, (3) region overlaps 5' end of gene, and (4) region overlaps 3'end of gene. *) human 1:many orthologs **) human many:many orthologs.

Table S4. The average overall heterozygosity and standard deviation for all fourteen breed groups.

References

1. Rubin C, Zody MC, Eriksson J, Meadows JRS, Sherwood E, *et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464: 587-591.
2. Darwin C (1868) The Variation of Animals and Plants under Domestication. Macmillan Publishers Limited.
3. Baker C (1968) Molecular genetics of avian proteins. IX. Interspecific and intraspecific variation of egg white proteins of the genus gallus. Genetics 58: 211-226.
4. Fumihito A, Miyake T, Sumi S, Takada M, Ohno S, *et al.* (1994) One subspecies of the red junglefowl (Gallus gallus gallus) suffices as the matriarchic ancestor of all domestic breeds. Proceedings of the National Academy of Sciences of the United States of America 91: 12505-12509.
5. Fumihito A, Miyake T, Takada M, Shingu R, Endo T, *et al.* (1996) Monophyletic origin and unique dispersal patterns of domestic fowls. Proceedings of the National Academy of Sciences of the United States of America 93: 6792-6795.
6. Hillel J, Groenen M, Tixier-Boichard M, Korol A, David L, *et al.* (2003) Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. Genetics Selection Evolution 35: 533 - 557.
7. Hutt FB (1949) Genetics of the Fowl. New York: McGraw Hill Book Company Inc.
8. Plant J (1986) The origin, evolution, history and distribution of the domestic fowl. Part 3. The Gallus species. Jungle fowls. 5 Bonar street, Maitland 2320, N.S.W., Australia: Privately published.
9. Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, *et al.* (2008) Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. PLoS Genetics 4: e1000010.

10. Liu Y, Wu G, Yao Y, Miao Y, Luikart G, *et al.* (2006) Multiple maternal origins of chickens: out of the Asian jungles. Molecular phylogenetics and evolution 38: 12 - 19.
11. Oka T, Ino Y, Nomura K, Kawashima S, Kuwayama T, *et al.* (2007) Analysis of mtDNA sequences shows Japanese native chickens have multiple origins. Animal Genetics 38: 287 - 293.
12. Kanginakudru S, Metta M, Jakati R, Nagaraju J (2008) Genetic evidence from Indian red jungle fowl corroborates multiple domestication of modern day chicken. BMC Evolutionary Biology 8: 174.
13. Crawford RD (1990) Poultry Breeding and Genetics. New York: Elsevier Science.
14. Havenstein GB, Ferket PR, Qureshi MA (2003) Growth, livability, and feed conversion of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. Poultry Science 82: 1500-1508.
15. Burt DW (2005) Chicken genome: Current status and future opportunities. Genome Research 15: 1692-1698.
16. Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genetics Research 23: 23-35.
17. Berry AJ, Ajioka JW, Kreitman M (1991) Lack of Polymorphism on the Drosophila Fourth Chromosome Resulting From Selection. Genetics 129: 1111-1117.
18. Zekarias B, Huurne AAHMT, Landman WJM, Rebel JMJ, Pol JMA, *et al.* (2002) Immunological basis of differences in disease resistance in the chicken. Veterinary Research 33: 109-125.
19. Julian RJ (1998) Rapid growth problems: ascites and skeletal deformities in broilers. Poulty Science 77: 1773-1780.
20. Whitehead C, Fleming R (2000) Osteoporosis in cage layers. Poultry Science 79: 1033-1041.
21. Balog JM (2003) Ascites Syndrome (Pulmonary Hypertension Syndrome) in Broiler Chickens: Are We Seeing the Light at the End of the Tunnel? Avian and Poultry Biology Reviews 14: 99 -126.
22. Baghbanzadeh A, Decuypere E (2008) Ascites syndrome in broilers: physiological and nutritional perspectives. Avian Pathology 37: 117 - 126.
23. Julian RJ (1993) Ascites in poultry. Avian Pathology 22: 419 - 454.
24. Rabie TSKM, Crooijmans RPMA, Bovenhuis H, Vereijken ALJ, Veenendaal T, *et al.* (2005) Genetic mapping of quantitative trait loci affecting susceptibility in chicken to develop pulmonary hypertension syndrome. Animal Genetics 36: 468-476.

25. Peiris BL, Ralph J, Lamont SJ, Dekkers JCM (2010) Predicting allele frequencies in DNA pools using high density SNP genotyping data. Animal Genetic. 42(1):113-6.
26. Yang H-C, Liang Y-J, Huang M-C, Li L-H, Lin C-H, *et al.* (2008) A genome-wide study of preferential amplification/hybridization in microarray-based pooled DNA experiments. Nucleic Acids Research 34: e106.
27. Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshere M, *et al.* (2000) Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. Human Genetics 107: 488-493.
28. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
29. Nei M (1972) Genetic distance between populations. American Naturalist 106: 283-292.
30. Megens HJ, Crooijmans RPMA, San Cristobal M, Hui X, Li N, *et al.* (2008) Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. Genetics Selection Evolution 40: 103–128.
31. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596-1599.
32. Eding H, Crooijmans RPMA, Groenen MA, Meuwissen THE (2002) Assessing the contribution of breeds to genetic diversity in conservation schemes. Genetics Selection Evolution 34: 613-633.
33. Muir WM, Wong GKS, Zhang Y, Wang J, Groenen MAM, *et al.* (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proceedings of the National Academy of Sciences of the United States of America 105 17312–17317.
34. Dana N, Megens HJ, Crooijmans RPMA, Olivier H, Mwacharo J, *et al.* (2010) East Asian contributions to Dutch traditional and western commercial chickens inferred from mtDNA analysis. Animal Genetics. In press.
35. Amills M, Jimenez N, Villalba D, Tor M, Molina E, *et al.* (2003) Identification of three single nucleotide polymorphisms in the chicken insulin-like growth factor 1 and 2 genes and their associations with growth and feeding traits. Poultry Science 82: 1485-1493.

36. Zhou H, Mitchell A, McMurtry J, Ashwell C, Lamont S (2005) Insulin-like growth factor-I gene polymorphism associations with growth, body composition, skeleton integrity, and metabolic traits in chickens. Poultry Science 84: 212-219.
37. Shimada M, Tritos NA, Lowell BB, Flier JS, Maratos-Flier E (1998) Mice lacking melanin-concentrating hormone are hypophagic and lean. Nature 396: 670-674.
38. Gerdin AK, Surve VV, Jönsson M, Bjursell M, Björkman M, *et al.* (2006) Phenotypic screening of hepatocyte nuclear factor (HNF) 4-gamma receptor knockout mice. Biochemical and Biophysical Research Communications 349: 825-832.
39. Ankra-Badu GA, Bihan-Duval EL, Mignon-Grasteau S, Pitel F, Beaumont C, *et al.* (2010) Mapping QTL for growth and shank traits in chickens divergently selected for high or low body weight. Animal Genetics 41: 400-405.
40. Desai J, Shannon ME, Johnson MD, Ruff DW, Hughes LA, *et al.* (2006) Nell1-deficient mice have reduced expression of extracellular matrix proteins causing cranial and vertebral defects. Hum an Molecular Genetics 15: 1329-1341.
41. Bokui N, Otani T, Igarashi K, Kaku J, Oda M, *et al.* (2008) Involvement of MAPK signaling molecules and Runx2 in the NELL1-induced osteoblastic differentiation. FEBS Letters 582: 365-371.
42. Zhou H, Deeb N, Evock-Clover CM, Mitchell AD, Ashwell CM, *et al.* (2007) Genome-Wide Linkage Analysis to Identify Chromosomal Regions Affecting Phenotypic Traits in the Chicken. III. Skeletal Integrity. Poultry Science 86: 255-266.
43. Hu Z-L, Fritz ER, Reecy JM (2007) AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. Nucleic Acids Research 35: D604–D609.
44. International Chicken Polymorphism Map Consortium (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432: 717-722.
45. Groenen MA, Wahlberg P, Foglio M, Cheng H, Megens H, *et al.* (2009) A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res 19: 510-519.

46. Megens H-J, Crooijmans R, Bastiaansen J, Kerstens H, Coster A, *et al.* (2009) Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genetics 10: 86.
47. Elferink M, van As P, Veenendaal T, Crooijmans R, Groenen M (2010) Regional differences in recombination hotspots between two chicken populations. BMC Genetics 11: 11.
48. Heldin C-H, Westermark B (1999) Mechanism of Action and In Vivo Role of Platelet-Derived Growth Factor. Physiological Reviews 79: 1283-1316.
49. Uutela M, Lauren J, Bergsten E, Li X, Horelli-Kuitunen N, *et al.* (2001) Chromosomal Location, Exon Structure, and Vascular Expression Patterns of the Human PDGFC and PDGFD Genes. Circulation 103: 2242-2247.
50. Ding H, Wu X, Kim I, Tam PPL, Koh GY, *et al.* (2000) The mouse Pdgfc gene: dynamic expression in embryonic tissues during organogenesis. Mechanisms of Development 96: 209-213.
51. Choi SJ, Marazita ML, Hart PS, Sulima PP, Field LL, *et al.* (2009) The PDGF-C regulatory region SNP rs28999109 decreases promoter transcriptional activity and is associated with CL/P. European Journal of Human Genetics 17: 774-784.
52. Melendrez CS, Meizel S (1995) Studies of porcine and human sperm suggesting a role for a sperm glycine receptor/Cl- channel in the zona pellucida-initiated acrosome reaction. Biology of Reproduction 53: 676-683.
53. Kumar P, Meizel S (2008) Identification and spatial distribution of glycine receptor subunits in human sperm. Reproduction 136: 387-390.
54. Rees MI, Lewis TM, Kwok JBJ, Mortier GR, Govaert P, *et al.* (2002) Hyperekplexia associated with compound heterozygote mutations in the beta-subunit of the human inhibitory glycine receptor (GLRB). Human Molecular Genetics 11: 853-860.
55. Kingsmore SF, Giros B, Suh D, Bieniarz M, Caron MG, *et al.* (1994) Glycine receptor beta-subunit gene mutation in spastic mouse associated with LINE-1 element insertion. Nature Genetics 7: 136-142.
56. Yonash N, Bacon LD, Witter RL, Cheng HH (1999) High resolution mapping and identification of new quantitative trait loci (QTL) affecting susceptibility to Marek's disease. Animal Genetics 30: 126-135.
57. de Koning DJ, Windsor D, Hocking PM, Burt DW, Law A, *et al.* (2003) Quantitative trait locus detection in commercial broiler lines using candidate regions. Journal of Animal Science 81: 1158-1165.

58. Buitenhuis AJ, Rodenburg TB, Siwek M, Cornelissen SJB, Nieuwland MGB, *et al.* (2004) Identification of QTLs Involved in Open-Field Behavior in Young and Adult Laying Hens. Behavior Genetics 34: 325-333.
59. De Koning DJ, Haley CS, Windsor D, Hocking PM, Griffin H, *et al.* (2004) Segregation of QTL for production traits in commercial meat-type chickens. Genetical Reseach 83: 211-220.
60. Navarro P, Visscher PM, Knott SA, Burt DW, Hocking PM, *et al.* (2005) Mapping of quantitative trait loci affecting organ weights and blood variables in a broiler layer cross. British Poultry Science 46: 430-442.
61. Sharman PWA, Morrice DR, Law AS, Burt DW, Hocking PM (2007) Quantitative trait loci for bone traits segregating independently of those for growth in an F2 broiler x layer cross. Cytogenetic and Genome Research 117: 296-304.
62. Rowe SJ, Pong-Wong R, Haley CS, Knott SA, De Koning DJ (2009) Detecting parent of origin and dominant QTL in a two-generation commercial poultry pedigree using variance component methodology. Genetics Selection Evolution Jan: 41:46.
63. Gao Y, Du ZQ, Wei WH, Yu XJ, Deng XM, *et al.* (2009) Mapping quantitative trait loci regulating chicken body composition traits. Animal Genetics 40: 952-954.
64. Ikeobi CON, Woolliams JA, Morrice DR, Law A, Windsor D, *et al.* (2004) Quantitative trait loci for meat yield and muscle distribution in a broiler layer cross. Livestock Production Science 87: 143-151.
65. Silverstein RA, Ekwall K (2005) Sin3: a flexible regulator of global gene expression and genome stability. Current Genetics 47: 1-17.
66. Dannenberg J-H, David G, Zhong S, van der Torre J, Wong WH, *et al.* (2005) mSin3A corepressor regulates diverse transcriptional networks governing normal and neoplastic growth and survival. Genes & Development 19: 1581-1595.
67. David G, Grandinetti KB, Finnerty PM, Simpson N, Chu GC, *et al.* (2008) Specific requirement of the chromatin modifier mSin3B in cell cycle exit and cellular differentiation. Proceedings of the National Academy of Sciences of the United States of America 105: 4168-4172.
68. Wu G, Lin Y-T, Wei R, Chen Y, Shan Z, *et al.* (2008) Hice1, a Novel Microtubule-Associated Protein Required for Maintenance of Spindle Integrity and Chromosomal Stability in Human Cells. Molecular and Cellular Biology 28: 3652-3662.

69. Holford KC, Edwards KA, Bendena WG, Tobe SS, Wang Z, *et al.* (2004) Purification and characterization of a mandibular organ protein from the American lobster, Homarus americanus: a putative farnesoic acid O-methyltransferase. Insect Biochemistry and Molecular Biology 34: 785-798.
70. Ikeobi CO, Woolliams JA, Morrice DR, Law A, Windsor D, *et al.* (2002) Quantitative trait loci affecting fatness in the chicken. Animal Genetics 33: 428-435.
71. Atzmon G, Blum S, Feldman M, Cahaner A, Lavi U, *et al.* (2008) QTLs Detected in a Multigenerational Resource Chicken Population. Journal of Heredity 99: 528-538.

6

General discussion

6.1 Introduction

For future improvements in production traits and animal welfare as well as to address future consumer demands, it is necessary to understand the etiology and biology underlying production traits and diseases. The primary aim of the research described in this thesis was to investigate the utility of several molecular approaches to identify causative variants underlying a variety of traits in the chicken. The identification of causative variants is a multistep process that is essentially similar for both monogenic and polygenic traits. The first step involves mapping to identify genomic regions that are associated with a particular trait or show evidence for selection. After the identification of these genomic regions, the next steps are the identification of all variants linked to these regions, and to obtain (biological relevant) evidence to identify the true causative variant(s).

Successful mapping is dependent on many different characteristics of the causative variants including the allele frequency, size of the phenotypic effect, evolutionary age, and the amount of selective pressure that the variant underwent (Figure 6.1). The phenotypic effect and allele frequency both influence the sample size needed to obtain sufficient statistical evidence for association. The evolutionary age of, and selective pressure on variants, as well as population demography, such as bottlenecks, expansion, admixture, inbreeding, and genetic drift, will affect the size of the haplotype in which the variant is located. The size of the haplotype influences the marker density needed for mapping. Due to the availability of reference genomes, high numbers of SNP markers, and high-throughput genotyping techniques, genome-wide assays with high marker densities became available in the last decade. This increased marker density lead to the development of genome-wide association (GWA) studies in addition to classical linkage mapping. Furthermore, the recent developments in massive parallel sequencing (MPS) technologies have increased the mapping resolution to a single base pair.

Successful detection of the true causative variant(s) relies on the mapping resolution, type of variants, knowledge on biological mechanisms involved in the trait, and on accurate gene annotation. If the mapping resolution is low, detection of causative variants becomes challenging due to a large numbers of possible candidate genes. Each type of variant such as SNP, small indel, copy number, and copy neutral variant, requires different methods for optimal detection. For instance, sequencing both alleles of the coding regions of *PRLR* and *SPEF2* would not have resulted in the detection of the CNV at the late feathering locus (Chapter 3). Detailed knowledge about the biological mechanisms involved in the trait under investigation, combined with accurate gene annotation permits more accurate

identification of candidate genes, even if the mapping resolution is low. However, accurate gene annotation in many species is often lacking, and gene functions are often based on orthologous genes in species such as human or mice. Predicting the gene function based on other species can introduce a bias. If the evolutionary distance is large, such as between chicken and human, divergent evolution might have resulted in different functions for the same ortholog.

The optimal strategy to identify causative variant(s) affecting a trait will be different for monogenic or polygenic traits. Moreover, these strategies will also depend on differences in population demography. In this final chapter, I discuss the implementation of currently available methods to detect causative variants with small and large phenotypic effects in monogenic or polygenic traits.

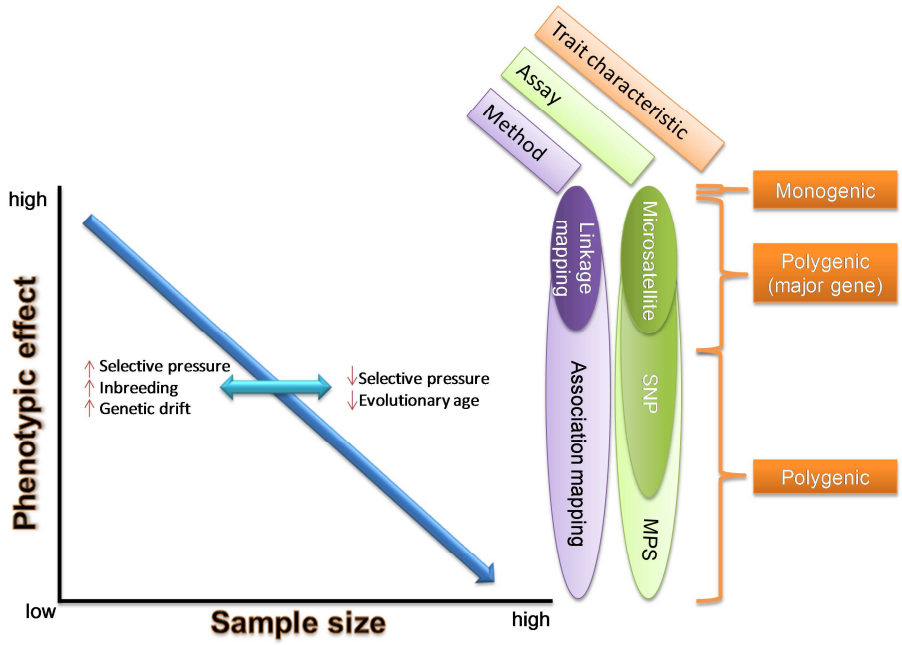


Figure 6.1 General overview of the strategies needed for causative variant mapping. Left part: the sample size needed for mapping will increase with decreasing phenotypic effect of variants. The sample size is moreover influenced by selective pressure on the variant, population demography such as inbreeding and genetic drift, and the evolutionary age of the variant. Right part: the different methods and assay types most appropriate for monogenic or polygenic traits.

6.2 Monogenic traits

A causative variant underlying a monogenic trait explains 100% of the phenotypic effect of the trait. As a result of this large effect, a small sample size and a low marker density are sufficient to successfully map the causative variant. However, linkage analysis with a low marker density will result in large confidence intervals, and subsequently a large number of candidate genes. The identification of causative variants will, therefore, heavily depend on the existence of clear functional candidate genes. For example, several studies successfully applied linkage analysis with small sample sizes and a limited number of markers, such as the double muscle phenotype in cattle [1] and the trait 'fishy taint of eggs' in chicken [2]. Both studies successfully detected the true causative variant of the trait because there was a clear functional candidate gene based on the function in human or mice. For the double muscle phenotype in cattle, knockout studies of a candidate gene, *MSTN*, resulted in a similar phenotype in mice. Likewise, for the 'fishy taint of eggs' trait, one candidate gene, *FMO3*, was known to be involved in trimethylaminuria, a disorder in human which results in odor reminiscent of rotting fish.

Nevertheless, for many linkage analysis studies obvious functional candidate genes are not immediately evident. Functional candidate genes may be difficult to identify because the biochemical pathways underlying the traits are not fully understood, or because gene annotation is incomplete. Sequencing all genes within the confidence intervals of associated regions is usually not achievable (at least before MPS) and, therefore, fine-mapping is needed to reduce the number of candidate genes. This fine-mapping includes, for instance, increased marker density in associated regions, LD mapping, backcrossing, advanced intercross lines, and identical-by-decent mapping [3]. Even after fine-mapping, extensive sequencing efforts are needed to detect the true causative variant within the remaining candidate genes. Although several examples exist that lead to easy identification of the causative variant, linkage mapping with low marker density is generally labor intensive due to the linkage mapping, fine-mapping, and extensive sequencing efforts needed.

The developments in genome-wide assays contribute to more powerful and efficient mapping of variants involved in different monogenic traits as shown by Charlier *et al.* [4]. Genome-wide SNP assays were successfully implemented to identify homozygous regions shared between affected individuals for five monogenic recessive traits in cattle. The benefit of the increased marker density is shown by the fact that for three of the five monogenic traits, linkage analysis and

association studies using a panel of 400 microsatellite markers did not result in successful mapping. Sequencing functional candidate genes within the associated regions lead to the identification of the causative variant in three out of the five monogenic traits. For all three monogenic traits in which the causative variant was detected, a clear relationship could be established between the gene function in human and the phenotypic characteristics of the trait in cattle, once again emphasizing that successful and quick identification of causative variants are aided by clear functional relationships between genes and traits. For one of the two traits, the crooked tail syndrome for which no causative variant was identified, none of the nineteen genes within the associated region had a function that could be related to the trait. Although this number of genes within the associated region is much less than typically found in linkage analysis studies, fine-mapping was still needed to reduce the number of candidate genes because sequencing all genes was labor intensive and expensive. A follow-up study using fine-mapping with additional markers and cases resulted in a region that comprised 7 genes [5]. Sequencing of the coding regions of these genes resulted in the identification of the causative variant underlying the crooked tail syndrome. Although genome-wide assays will increase the mapping resolution, identification of the causative variant still requires additional fine-mapping and sequencing efforts, especially when there are no obvious candidate genes.

Recent developments show that massive parallel sequencing (MPS) is likely going to replace the genome-wide assays in future studies that aim to detect causative variants underlying monogenic traits. One major benefit for MPS based strategies is the power to detect the causative variant in a single study, which eliminates fine mapping and sequencing of candidate genes. The power of MPS to detect causative variants in monogenic traits based on a small number of affected individuals has recently been demonstrated in humans (see box 2 in Chapter 1 for more details). Monogenic traits are often caused by coding variants and, therefore, the earliest of these studies applied targeted re-sequencing of the exome to detect coding SNPs (e.g. [6,7,8,9]). This targeted sequence capture was needed to increase the sequence coverage per base - and therefore the reliability of variant calling – because adequate coverage in whole-genome re-sequencing was too expensive. The costs of MPS are rapidly decreasing and, therefore, it is expected that costs for whole-genome re-sequencing will become similar to that for exome sequencing combined with sequence capture. The difference in costs between whole-genome and exome sequencing in livestock species is momentarily already small because exome capture arrays or assays are not commercially available and, therefore, more expensive. In addition, exome sequencing is only targeted to coding regions,

and is less usefull for copy number variation (CNV) detection because of biases in the read depth due to the capturing steps. Whole-genome re-sequencing for chicken is cost-effective due to the small size of the genome and provides an increased coverage per base compared to, for instance, human (3 fold increase). In the remaining part of this discussion I will refer to whole-genome sequencing for MPS based sequencing strategies in the chicken.

Despite the decreasing costs, whole-genome sequencing is still expensive compared to genome-wide assays. Recently, Sobreira *et al.* [10] successfully detected the causative variant involved in a monogenic trait in human by combining data from a linkage analysis based on genome-wide assays with sequencing data of a single affected individual. In essence, the single individual was re-sequenced to identify all possible causative variants (109 protein affecting variants in total), and linkage analysis was used for fine-mapping (1 protein affecting variant in the associated regions). A similar approach is currently feasible and cost-effective in the chicken. Nevertheless, MPS costs will further decrease in the coming years and therefore the preferred method will become whole-genome re-sequencing of families with affected and unaffected individuals, or unrelated individuals with opposite phenotypes for the monogenic trait.

All the studies described above were focused on the identification of (coding) SNPs or small indels. Structural variants (SV), such as CNVs (insertions and deletions) and copy neutral variants (inversions and translocations), are increasingly being recognized to affect phenotypic variation. One example of a CNV has been described in chapter 3 where we describe the identification of a 180 kb tandem duplication as the causative variant for the late feathering phenotype. Another example of a SV that underlies a monogenic trait is the 3.2 kb CNV located within the gene S*ox5* that causes the Pea comb phenotype in chicken [11]. Therefore, future studies that aim to detect causative variants need to include SV identification. Array-based comparative genomic hybridization (aCGH) is a method that enables genome-wide CNV detection. Recently, two studies were performed in the chicken that used this method [12] [13]. Although both studies detected the large tandem duplication in the late feathering phenotype, both failed to detect the small CNV in *Sox5* due to insufficient resolution. Moreover, with aCGH it is not possible to determine the exact number of copies within a CNV, nor is it possible to determine the genomic organization of the SV. In addition, copy neutral variants such as inversions and translocations can also not be detected with aCGH. Although the intensity signals of SNPs in genome-wide SNP assays can also be used for the detection of CNVs, these assays suffer from the same shortcoming as aCGH. Furthermore, SNP assays suffer from ascertainment bias, particularly for SNPs

located within SV regions. SNPs located in SV regions are often excluded from the assay due to non-Mendelian inheritance. The shortcomings of aCGH and SNP assays for the detection of SV can be circumvented by MPS. Read depth analysis provides exact copy numbers [14,15], split reads analysis provides information with regard to the location of breakpoint(s) [16], and paired-end mapping allows the detection of both copy number and copy neutral variants [17,18]. Moreover, the resolution is much higher compared to either aCGH or SNP assays. Nevertheless, even with MPS based strategies reconstruction of complex SV remains challenging.
In summary, existing methods and techniques are currently available to perform cost effective studies to detect causative variants underlying monogenic traits. At the moment, combining the relatively cheap genome-wide assays with expensive MPS will be the preferred method for cost-effective yet powerful detection of causative variants. However, if the costs of sequencing will decrease in the near future, the preferred method for future studies will be solely based on MPS.

6.3 Polygenic traits

Although strategies to detect causative variants underlying monogenic trait are currently powerful and affordable, most important production traits and genetic diseases in livestock species have a polygenic background [19]. Here I will address the implementation of current methods to detect causative variants underlying polygenic traits, with a distinction between polygenic traits with or without the presence of (a) major gene(s).

Polygenic traits with major genes

Major genes are variants that explain a substantial part of the phenotypic variance of a trait and are nearly following the pattern of Mendelian inheritance [20]. Therefore, the strategy used to detect causative variants in major genes is similar to the strategy used for monogenic traits. However, the genotype - phenotype relation is not absolute and, therefore, larger sample sizes are needed for association. The additional genotyping, phenotyping, and data-analysis required results in more expensive studies compared to those in monogenic traits.
In livestock species, there are several successes in the identification of causative variants in major genes with strategies similar to those used for monogenic traits. Examples are the missense variant in *DGAT1* involved in milk yield and composition [21], a regulatory variant in *IGF2* involved in muscle growth in pigs [22], and a regulatory variant in *GDF8* involved in muscularity in sheep [23]. For all these studies, a genome-wide linkage analysis was performed with a few hundred microsatellite markers, subsequently followed by fine-mapping and candidate gene

sequencing. While most detected variants in monogenic traits are located in the coding regions, two out of the three causative variants within the major genes are in the regulatory part of the gene. In general, such variants will require more sophisticated, time consuming, and expensive follow-up studies to determine the true causative variant.

Several studies have suggested the involvement of one or two major genes in the pulmonary hypertension syndrome (PHS) [24,25,26]. We have described a large experiment aimed at the genetic characterization of PHS and in our study we do not have any indications for the involvement of major genes in PHS (Chapter 4). Instead, our results show that PHS is influenced by large number of genes each with small phenotypic effects. This is in agreement with the results of a previous linkage mapping study [27] and recently performed GWA study (personal communication R. Okimoto, Cobb-Vantress Inc.). A recent study suggests that at least 100k SNPs are needed to capture the majority of haplotypes within a broiler population [28]. Both GWA studies have only included a fraction of this number and, therefore, it cannot be excluded that (major) genes are undetected due to insufficient marker resolution (resulting in no or insufficient LD between assayed markers and the causative variant). In addition, for the gaps and missing microchromosomes in the reference genome [29], markers are lacking in the assays, resulting in additional uncovered regions of the genome. For example, *NOS3* (endothelial nitric oxide synthase) has been suggested to be involved in PHS susceptibility [30]. Although it is known that eNOS is present in chicken - *eNOS* mRNA is expressed in chicken - the gene it not present in the current genome build. Because *eNOS* is missing from the reference genome of chicken, markers within this gene are not included in the genome-wide assay.

The uncovered regions in GWA studies, especially the missing microchromosomes, emphasize the need to improve the current reference genome. Improvements are also essential for MPS based strategies because short read alignment currently depends on the reference genome. This dependency will remain until accurate *de novo* assemblies based on MPS become possible. Although several methods are developed for *de novo* assembly of MPS data, the most recent *de novo* assemblies of the human genome are not very accurate [31]. Alkan *et al.* [31] compared two *de novo* assemblies of a human genome (Li 2010) to the human reference genome and detected that the *de novo* assemblies were 16.2% shorter, and that 420 Mb of common repeats, 99.1% of verified segmental duplication, and 2,377 coding exon were missing. In the genome-wide SNP assay used to construct the linkage map described in chapter 2, we included several hundred markers that had previously not been mapped to the reference genome. We anticipated that these markers had

a high likelihood of being located on one of the missing microchromosomes or in sequence gaps in the reference genome. The majority of informative markers were successfully mapped to gaps in the reference genome, thereby facilitating improvements that can be used in future genome builds. However, no new linkage groups were detected that represented one of the missing microchromosomes. New MPS efforts were recently undertaken to close these gaps in the reference genome [32]. The study resulted in successful mapping of additional sequences to gaps in known chromosomes. Moreover, two new linkage groups were identified that might represent missing microchromosomes. Nevertheless, several microchromosomes are still missing or underrepresented in the improved reference genome. New efforts to close the remaining gaps should be attempted in the near future. It is not clear why the sequences of the several microchromosomes are completely missing in the current draft sequence of the chicken genome. One assumption is that the sequences are missing because they are difficult to clone and propagate in *E.coli [33]*. If this cloning step is indeed the problem, MPS based strategies should have been more successful which, apart from the two new linkage groups identified, appears not to be the case. It is, therefore, likely that the sequencing of the missing microchromosomes will not be solved with current second generation MPS technologies alone. Other explanations for the difficulties in sequencing these microchromosomes are the high GC nucleotide contents [29] that results in PCR amplication difficulties. Third generation sequencing techniques provide a solution if PCR amplification is difficult because single molecule sequencing does not include PCR amplification steps. The assembly of the microchromosomes might be assisted by techniques such as optical mapping [34,35,36]. In optical mapping it is possible to obtain the restriction pattern for single chromosomes. This restriction pattern can assist in accurate assembly of contigs found for the new chromosomes. However, because contigs need to include multiple restriction sites for accurate mapping the size of these contigs need to be large. Large contigs are generally not feasible with second generation MPS and, therefore, require third generation sequencing.

Polygenic traits without major genes

While the detection of causative variants underlying monogenic and polygenic traits with major genes have seen a number of successes, the detection of variants involved in polygenic traits has proven to be extremely challenging. For instance, less than 1% of 2,000 identified QTLs have been characterized at a molecular level in crosses of inbred strains of mice and rats, and almost all variants had large phenotypic effects [37]. Polygenic traits influenced by many variants with small

phenotypic effects do not follow Mendelian inheritance and, therefore, linkage analysis is excluded as a method to map variants. Instead, (genome-wide) association mapping is needed. The small phenotypic effect of these variants is difficult to distinguish from neutral variants and therefore requires large sample sizes for sufficient statistical power. Moreover, the selective pressure on each of the variants is typically low and the variants are often located in small haplotypes. In order to detect these haplotypes a high resolution assay is needed.

GWA studies in human indicated that even high resolution SNP assays combined with huge sample sizes may not be sufficient to detect all variants involved in polygenic traits. For example, a study on height in humans based on 180,000 individuals with 2.8 million (common) SNPs, resulted in the identified of hundreds of genetic variants each with a very small phenotypic effect [38]. Despite this large number of variants identified, all variants combined explained only 10% of the phenotypic variation of human height. Because it is estimated that approximately 80% of the total variation in human height is attributed to additive genetic factors [39], a large part of the heritability of human height remains undetected. This missing heritability might, at least partially, be explained by (rare) causative variants that could not be captured by the common SNPs used for the genome-wide assays. This suggests that future association studies for some polygenic traits will require whole-genome re-sequencing of a large number of phenotyped individuals. Although technically feasible, the extreme high costs for sequencing, data-analysis, and phenotyping make such studies unrealistic.

One possibility to achieve cost-effective yet powerful studies is to combine the power of MPS and the affordability of genome-wide assays in genotype imputation. Genotype imputation refers to the prediction of missing genotypes in individuals genotyped at a relative low resolution, based on genotype information of a reference population in which individuals are genotyped at high resolution, for instance by whole-genome re-sequencing [40,41,42,43]. The central thought behind genotype imputation is that genotyping or sequencing a small part of the population (the so called reference population) results in the detection of the majority of haplotypes segregating within the entire population. The haplotype information in the reference population can, subsequently, be used to predict haplotypes in individuals that are genotyped at a lower resolution (the sample population) (Figure 6.2). Recently, a large number of genomes was sequenced to establish such a reference population for human [42] and illustrated the power of imputation for existing GWA studies. Near the original signal of the GWA study, several imputed genotypes had a substantial higher association signal. In addition, the size of the associated region could be reduced because only a subset of

imputed genotypes showed these high signals. Thus, genotype imputation results in increased power to detect variants, and enables further fine-mapping of the associated regions. Although current imputation methods are limited to SNP imputation, it is expected that upcoming methods will also be capable to impute indels and SV.

Accurate genotype imputation relies on two factors. The first factor is the number of individuals needed within the reference population to capture the majority of haplotypes segregating within a population. Haplotypes that are not detected in the reference populations are inaccurately imputed in the sample population. The second factor is the marker density needed for the genome-wide assay needed to genotype the sample population. If the design of these assays is not optimal, for instance because haplotypes detected within the reference population are not sufficiently covered by assayed markers, genotype imputation in the sample population will be inaccurate.

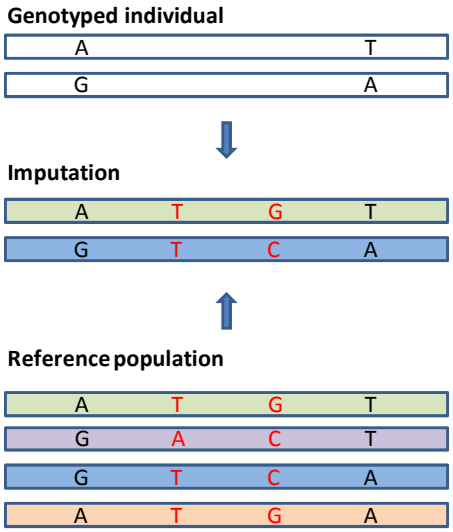


Figure 6.2 Genotype imputation. The figure illustrated a simplified example of genotype imputation within an individual genotyped for two markers (black nucleotide). Based on the haplotypes in the reference population genotypes can be imputed (red nucleotides) in the individual.

The optimal number of individuals needed for reference populations in livestock species depends on the population demography (LD structure and effective population size), the pedigree structure, and the desired imputation error rate [44]. Except for the 1000 genome project in human, there are no studies published that used whole-genome re-sequencing data for genotype imputation. In layer populations, whole-genome re-sequencing studies are planned that are essentially based on pedigree structure only (A. Vereijken, personal communication). For the layer populations the strategy is to sequence the genome of all sires of a particular generation (25 individuals), and the fathers of all dams within this generation (another 25 individuals). Thus, by sequencing 50 individuals, at least 75% of all haplotypes within the population will be captured by this approach. For commercial broiler populations, a similar strategy requires whole-genome re-sequencing of 80 individuals (A. Vereijken, personal communication).

Although the data described in chapter 4 provide an excellent opportunity for genotype imputation, it is expected that only a fraction of all haplotypes within the complete population are captured by sequencing the genomes of twelve individuals. Moreover, due to the extent of LD within broilers it is not certain that the 18k SNP assay used (effectively 10k segregating markers within the PHS population) will be sufficient for accurate imputation. It has previously been suggested that genome-wide assays should contain at least 100k markers to capture all haplotypes within a broiler population [28]. Therefore, it is expected that a large part of the haplotypes within the population will not be properly imputed, even if the reference population is large.

To examine potential shortcomings for genotype imputation based on the 18k SNP assay I focused on the haplotypes predicted within a 155 kb LD block on chromosome 27 (Figure 6.3). For this region, haplotypes were predicted in three datasets using Haploview [45]. The first dataset includes all 1,313 animals genotyped with the 18k SNP assay. The second data set includes a subset of this population, namely the 12 individuals that have been selected for whole-genome re-sequencing. The third dataset involved the genotypes for these 12 animals obtained with MPS (Chapter 4). Within the haploblock, based on the SNP genotypes of all 1,313 individuals, 10 haplotypes were predicted within the entire population. Only 40% of these haplotypes (4 out of 10) were detected in the twelve animals (genotyped with the 18k SNP assay) and, therefore, it is obvious that larger sample sizes are needed to capture all haplotypes within the population. Nevertheless, the 4 haplotypes present in the 12 animals represent 93.3% of the haplotype diversity present in the entire population, which indicates that for the majority of animals within this population accurate haplotypes can be imputed

based on the twelve animals. However, high resolution genotyping with MPS revealed that 12 haplotypes are actually underlying the 4 haplotypes predicted based on the 18k SNP assay. Thus, even for a region that is in high LD, the 18k SNP assay is inadequate to capture all underlying true haplotypes. For regions in the genome with low LD, it is expected that even much more haplotypes will be missed.

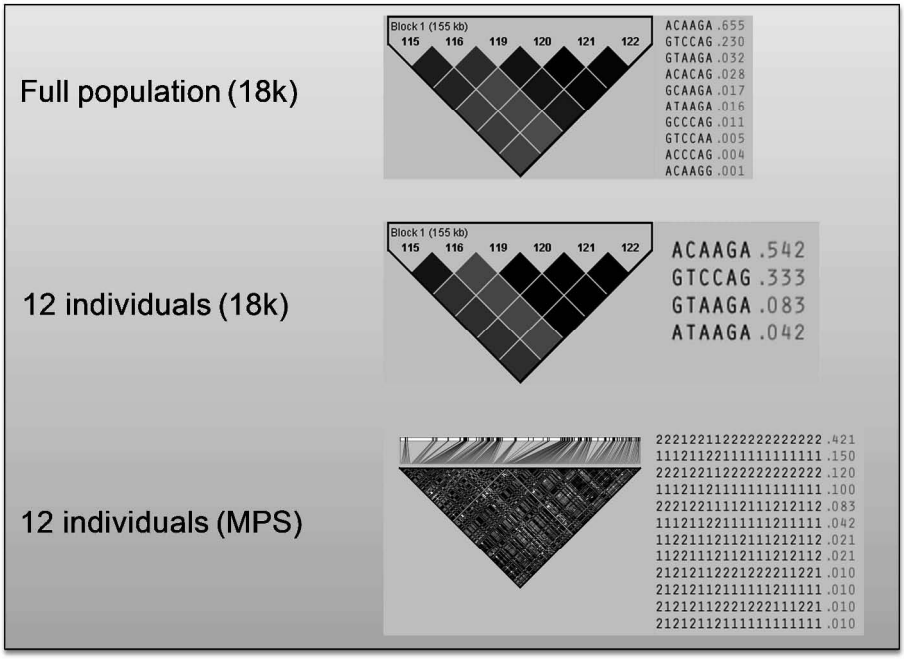


Figure 6.3 Haplotype structure of the full population genotyped with 18k SNPs, and the haplotype structure of twelve animals based on either the 18k or MPS data. In the figure the LD structure of the haploblock is shown (left), and the underlying haplotypes with their corresponding frequency (right). The high LD region is located on chromosome 27 (3,880,581-4,036,439 bp). All animals were considered unrelated in the Haploview analysis. Minor allele frequency per SNP > 0.05%. Note that for the MPS data haplotypes only the last few bases for the haplotype are shown, and that the bases are annotated 1 or 2 instead of the true nucleotides. The 12 animals genotyped with 18k are the same individuals as selected for whole-genome re-sequencing.

It is likely that the twelve sequenced genomes will not be particularly usefull to impute genotypes for individuals genotyped with the 18k SNP assay in the entire population. In the future, a better strategy for genotype imputation will be to first sequence the genomes of the reference population, subsequently followed by the development of a low resolution genome-wide assay in which SNPs should be selected to cover the majority of haplotypes within the genome. These 'low resolution' genome-wide assays nevertheless need to contain up 10-20k SNPs for a single layer population and 100k SNPs for a single broiler population as previously suggested [28].

6.4 Hitch-hiking mapping

Besides association based strategies, we applied a hitch-hiking mapping approach to identify regions in the chicken genome that are or have been under selection (sweep) (Chapter 5). The assumption underlying hitch-hiking mapping is that regions in the genome under selection must contain important functional variants. A well-known example is the selective sweep detected at the Lactase gene in human [46]. A variant located in the regulatory region of lactase gene (*LCT*) is involved in lactase persistence at adulthood [47]. This variant allowed adults to consume nutrition's from dairy animals and, therefore, underwent rapid positive selection [48]. In the chicken, a recently performed hitch-hiking study based on MPS identified a non-synonymous variant in the *TSHR* gene that might be involved in the absence of strict regulation of seasonal reproduction observed in domestic chickens [49] (Chapter 1, box 2).

Compared to association studies, hitch-hiking mapping does not require expensive phenotyping as it is based on exploiting the nucleotide variation within the genome. Instead, other criteria such as breeding goals or population characteristics can be used for initial interpretation of the function of the causative variant located in the sweeps. Unfortunately, these criteria are often rather vague, and include many possible pathways in which candidate genes might be involved. The identification of the true causative variant therefore remains a difficult task, especially when the size of the sweep is large and contains many genes. Because there is little *a priori* knowledge on the function of the gene under selection, each gene has to be considered as a potential candidate. A more sophisticated approach for hitch-hiking mapping is the use of selection lines for a particular trait, as recently shown by Burke *et al.* [50]. In this study, several genomic regions in Drosophila were identified that show strong allele frequency differentiation between a control population and a population selected for accelerated development. The identified regions are therefore instant target regions for the

trait under selection. The results of Burke *et al.* show that in specifically designed selection lines there is no need to phenotype individuals and to perform phenotypic based mapping studies. Disadvantages of this method are that the selection does not necessarily affect all variants underlying a trait, and that in particular variants with large phenotypic effects are identified. Moreover, selection lines are expensive and the design will take considerable time.

In chapter 5, we discussed that the successful detection of regions under selection relies on both genotype resolution and population demography. The marker density used in this study was sufficient to detect haplotypes that underwent recent and strong directional selection. Nevertheless, a higher marker resolution is needed detect 'ancient' sweeps, because these sweeps are generally smaller in size due to recombination events that occurred after the initial sweep. Therefore, it is desirable to use MPS in future hitch-hiking mapping studies. Another advantage for MPS based hitch-hiking mapping is that this strategy removes the ascertainment bias of SNP selection in genome-wide assays [51]. Successful implementation of hitch-hiking mapping also depends on population demography. In populations with low nucleotide diversity it is difficult to distinguish true signals of selection from signals that result from genetic drift or a population bottleneck. The commercial white egg-layers, for instance, have many shared homozygous regions within the genome due to a major bottleneck that occurred when the lines were established from a small base population [52]. In order to distinguish true sweeps from genetic drift, additional information is needed from other breeds that share the same breeding goal, for instance egg production. If the same region is identified in both breeds, there is more confidence that the region is truly under selection. Because such studies need a relatively high number of individuals, hitch-hiking mapping studies based on MPS are still costly. To reduce costs, pooling the DNA of individuals per population provides a cost-effective approach [49]. The disadvantage of DNA pooling, however, is that haplotype information is lost. Haplotype information is required to detect incomplete sweeps and balancing selection [53]. With the increased throughput of sequencing platforms, tag-based pooling of individuals provides a good and less expensive alternative compared to DNA pooling, while retaining haplotype information [54].

6.5 Final thoughts

Technological developments in the last decade have led to a true revolution in the field of genetics. Although many challenges lie ahead, strategies are available to successfully detect causative variants underlying monogenic or polygenic traits. The technologies for these strategies are currently available, although the implementation is often still limited by the size of the available budget. Nevertheless, combining different methods can result in cost-effective yet powerful studies. The costs for whole-genome re-sequencing are expected to reduce further in the near future, making it feasible to include more individuals and boost the power of association studies. However, increasingly the bottle-neck in association studies is the accuracy of the phenotypes available. Therefore, one has to consider whether it wouldn't be better to spend the available budget to increase the power of association studies through more accurate phenotyping instead of increasing the sample sizes or marker densities.

The recent technological developments contribute to more accurate mapping of traits. Subsequently, the future challenge will become the identification of the true (biological relevant) causative variants underlying the traits. In order to detect these causative variants it is essential to have a high quality reference genome with accurate gene annotation. For most livestock species, improvements are still needed in both areas. Fortunately, decreasing cost of MPS enables studies to improve gene annotation with methods such as RNA-seq [55]. For instance, whole transcriptome sequencing for many different tissues is currently affordable in all species.

With the third generation sequencing technologies in sight, it is expected that the necessary improvements will be addressed in the near future. At the same time, it is important to realize that the developments of these third generation sequencing technologies will results in significant (computational) challenges in the near future. To handle the massive amounts of data that will be provided by these new sequencing methods, development in software and hardware are required, as well as researchers trained in bioinformatics.

References

1. Grobet L, Royo Martin LJ, Poncelet D, Pirottin D, Brouwers B, *et al.* (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. Nat Genet 17: 71-74.
2. Honkatukia M, Reese K, Preisinger R, Tuiskula-Haavisto M, Weigend S, *et al.* (2005) Fishy taint in chicken eggs is associated with a substitution within a conserved motif of the FMO3 gene. Genomics 86: 225-232.
3. Hu X, Gao Y, Feng C, Liu Q, Wang X, *et al.* (2009) Advanced technologies for genomic analysis in farm animals and its application for QTL mapping. Genetica 136: 371-386.
4. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, *et al.* (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat. Genet. 40: 449-454.
5. Fasquelle C, Sartelet A, Li W, Dive M, Tamma N, *et al.* (2009) Balancing Selection of a Frame-Shift Mutation in the MRC2 Gene Accounts for the Outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. PLoS Genet 5: e1000666.
6. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.
7. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences 106: 19096-19101.
8. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, *et al.* (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet. 42: 790-793.
9. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. 42: 30-35.
10. Sobreira NLM, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, *et al.* (2010) Whole-Genome Sequencing of a Single Proband Together with Linkage Analysis Identifies a Mendelian Disease Gene. PLoS Genet 6: e1000991.
11. Wright D, Boije H, Meadows JRS, Bed'hom B, Gourichon D, *et al.* (2009) Copy Number Variation in Intron 1 of SOX5 Causes the Pea-comb Phenotype in Chickens. PLoS Genet 5: e1000512.

12. Wang X, Nahashon S, Feaster T, Bohannon-Stewart A, Adefope N (2010) An initial map of chromosomal segmental copy number variations in the chicken. BMC Genomics 11: 351.
13. Crooijmans R, Fife M, Fitzgerald T, Schmidt C, Cheng H, *et al.* (2011) Global variation in copy number in the chicken genome. *Submitted for publication*.
14. Alkan C, Kidd J, Marques-Bonet T, Aksay G, Antonacci F, *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061 - 1067.
15. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, *et al.* (2010) Diversity of Human Copy Number Variation and Multicopy Genes. Science 330: 641-646.
16. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Research 20: 623-635.
17. Korbel J, Urban A, Affourtit J, Godwin B, Grubert F, *et al.* (2007) Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318: 420 - 426.
18. Chen K, Wallis J, McLellan M, Larson D, Kalicki J, *et al.* (2009) Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6: 677 - 681.
19. Andersson L (2001) Genetic dissection of phenotypic diversity in farm animals. Nature Rev Genet 2: 130-138.
20. Singleton AB, Hardy J, Traynor BJ, Houlden H (2010) Towards a complete resolution of the genetic architecture of disease. Trends in Genetics 26: 438-442.
21. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, *et al.* (2002) Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. Genome Research 12: 222-231.
22. Van Laere A-S, Nguyen M, Braunschweig M, Nezer C, Collette C, *et al.* (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425: 832-836.
23. Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, *et al.* (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat. Genet. 38: 813-818.

24. Navarro P, Visscher PM, Chatziplis D, Koerhuis ANM, Haley CS (2006) Segregation analysis of blood oxygen saturation in broilers suggests a major gene influence on ascites. British Poultry Science 47: 671 - 684.
25. Druyan S, Ben-David A, Cahaner A (2007) Development of Ascites-Resistant and Ascites-Susceptible Broiler Lines. Poult Sci 86: 811-822.
26. Druyan S, Cahaner A (2007) Segregation Among Test-Cross Progeny Suggests That Two Complementary Dominant Genes Explain the Difference Between Ascites-Resistant and Ascites-Susceptible Broiler Lines. Poult Sci 86: 2295-2300.
27. Rabie T, Crooijmans R, Bovenhuis H, Vereijken A, Veenendaal T, *et al.* (2005) Genetic mapping of quantitative trait loci affecting susceptibility in chicken to develop pulmonary hypertension syndrome. Animal Genetics 36: 468 - 476.
28. Megens H-J, Crooijmans R, Bastiaansen J, Kerstens H, Coster A, *et al.* (2009) Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genetics 10: 86.
29. Hillier L, Miller W, Birney E, Warren W, Hardison R, *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695 - 716.
30. Hamal KR, Wideman RF, Anthony NB, Erf GF (2010) Differential expression of vasoactive mediators in microparticle-challenged lungs of chickens that differ in susceptibility to pulmonary arterial hypertension. American Journal of Physiology - Regulatory, Integrative and Comparative Physiology 298: R235-R242.
31. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat. Methods. 8: 61-65.
32. Groenen M, Megens H, Zare Y, Warren W, Hillier L, *et al.* (2011) Designing SNP chips: obstacles and lessons learned from development and characterization of a 60K chip for chicken. *Submitted for publication*.
33. Groenen M, Wahlberg P, Foglio M, Cheng H, Megens H, *et al.* (2009) A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res 19: 510 - 519.
34. Cai W, Jing J, Irvin B, Ohler L, Rose E, *et al.* (1998) High-resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. Proceedings of the National Academy of Sciences of the United States of America 95: 3390-3395.

35. Lin J, Qi R, Aston C, Jing J, Anantharaman TS, *et al.* (1999) Whole-Genome Shotgun Optical Mapping of Deinococcus radiodurans. Science 285: 1558-1562.
36. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56-64.
37. Flint J, Valdar W, Shifman S, Mott R (2005) Strategies for mapping and cloning quantitative trait genes in rodents. Nat. Rev. Genet. 6: 271-286.
38. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-838.
39. Visscher PM (2008) Sizing up human height variation. Nat. Genet. 40: 489-490.
40. Howie BN, Donnelly P, Marchini J (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5: e1000529.
41. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11: 499-511.
42. 1000 Genomes Project Consortium et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
43. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype Imputation. Annual Review of Genomics and Human Genetics 10: 387-406.
44. Druet T, Schrooten C, de Roos APW (2010) Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science 93: 5443-5454.
45. Barrett J, Fry B, Maller J, Daly M (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263 - 265.
46. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, *et al.* (2004) Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. The American Journal of Human Genetics 74: 1111-1120.
47. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, *et al.* (2002) Identification of a variant associated with adult-type hypolactasia. Nat. Genet. 30: 233-237.
48. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, *et al.* (2005) Genomic scans for selective sweeps using SNP data. Genome Research 15: 1566-1575.

49. Rubin C, Zody MC, Eriksson J, Meadows JRS, Sherwood E, *et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464: 587-591.
50. Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, *et al.* (2010) Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature 467: 587-590.
51. Pérez-Enciso M, Ferretti L (2010) Massive parallel sequencing in animal genetics: wherefroms and wheretos. Animal Genetics 41: 561-569.
52. Muir W, Wong G, Zhang Y, Wang J, Groenen M, *et al.* (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proc Natl Acad Sci USA 105: 17312 - 17317.
53. Pavlidis P, Hutter S, Stephan W (2008) A population genomic approach to map recent positive selection in model species. Molecular Ecology 17: 3585-3598.
54. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods. 5: 235-237.
55. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 5: 621-628.

Summary

Summary

The chicken currently provides more than a quarter of the meat and nearly all eggs produced worldwide. For future improvements in production traits and animal welfare as well as to address future consumer demands it is necessary to understand the etiology and biology underlying production traits and diseases. The primary aim of the research described in this thesis was to investigate the utility of several molecular approaches to identify causative variants underlying a variety of traits in the chicken.

The general introduction in **chapter 1** provides an overview of the domestication history of the chicken - with a particular focus on commercial chicken breeds - and describes the importance to identify causative variants underlying production traits and diseases. Furthermore, several different molecular techniques and methods are introduced that are being used to detect causative variants underlying monogenic and polygenic traits.

Linkage maps are essential for linkage analysis, important to study recombination rates and recombination hotspots within the genome and can assist in the sequence assembly of genomes. In **chapter 2** we describe the construction of a new high-resolution linkage map of the chicken genome based on two chicken populations with a total of 1619 individuals. The two populations used are a purebred broiler line and a broiler x broiler cross. This high resolution allowed accurate identification of recombination hotspots in the chicken genome, including sex specific recombination. Furthermore, to improve the current reference genome (WASHUC2), 613 unmapped markers were included in the genome-wide assay that included a total of 17,790 SNPs. The resulting linkage map comprises 13,340 SNPs, of which 360 had not been assigned to a known chromosome on chicken genome build WASHUC2. The resulting linkage map is composed of 31 linkage groups, with a total length of 3,054 cM for the sex-average map of the combined population. Regional differences in recombination hotspots between the two mapping populations were observed for several chromosomes near the telomere of the p arm. The sex-specific analysis revealed that these regional differences were mainly caused by female-specific recombination hotspots in the broiler × broiler cross.

In **chapter 3** we describe the molecular characterization of the locus causing the late feathering phenotype; a monogenic trait in chicken that results in a delayed emergence of flight feathers at hatch. The late feathering phenotype is beneficial to breeders as it can be used for sex typing at hatch. The locus has, therefore, been extensively used in diverse commercial chicken breeds. However, a retrovirus closely linked to the late feathering allele causes a negative pleiotropic effect on

egg production and causes viral infections. Within this chapter we describe the identification of a 180 kb tandem duplication in the late feathering allele using a quantitative PCR approach. The tandem duplication results in the partial duplication of two genes; the prolactin receptor and the gene encoding sperm flagellar protein 2. Sequence analysis revealed that the duplication is identical in broiler, white egg-layer, and brown egg-layer lines. This information was also used to design a molecular test to detect this duplication, particularly in heterozygous individuals.

The recent advances in massive parallel sequencing technologies have enabled rapid and cost-effective detection of all genetic variants within genomes. The detection of all genetic variants within a genome has further increased our ability to identify causative variants underlying quantitative trait loci (QTL). In **chapter 4**, we combined a genome-wide association study with whole-genome resequencing to identify causative variants underlying the pulmonary hypertension syndrome (PHS), a polygenic trait in chicken. PHS is a metabolic disease that has been linked to intense selection on growth rate and feed conversion ratio of modern broilers (meat-type chicken). PHS has become one of the most frequent causes of mortality within the broiler industry and leads to substantial economic losses and reduced animal welfare. In total, 18 QTL regions were identified in the genome-wide association study. In order to detect causative variants underlying these QTL regions, we sequenced the genomes of twelve individuals. To maximize the detection of causative variants we selected the individuals based on extreme phenotypes for PHS. Within 8 QTL regions we identified a total of 10 genes that contain at least one variant that is predicted to affect protein function. Moreover, 7.62 million SNPs were detected within the twelve animals compared to the reference genome. These markers can be used in the development of future genome-wide assays.

Genomic regions that have undergone selection should contain loci that influence important phenotypic traits and will, therefore, include causative variant(s) that could aid in further future improvement of production traits and disease resistance. In **chapter 5**, we applied hitch-hiking mapping to make a broad assessment of the effects of selection histories in domesticated chicken. Towards this end, we sampled commercial chickens representing all major breeding goals from multiple breeding companies. In addition, we sampled non-commercial chicken diversity by sampling almost all recognized traditional Dutch breeds and a representative sample of breeds from China. The broad sample of 67 commercial and non-commercial breeds were assessed for signatures of selection in the genome using information of 57,636 SNPs that were genotyped on pooled DNA samples. Our

approach demonstrates the strength of including many different populations with similar, and breed groups with different selection histories to reduce stochastic effects based on single populations. The detection of regions of putative selection resulted in the identification of several candidate genes that could aid in further improvement of production traits and disease resistance.

Finally, the general discussion **in chapter 6** describes the main findings of this thesis. In this chapter recommendations are given for the best strategies to detect causative variants underlying monogenic or polygenic traits. All strategies can benefit substantially from the recent developments in massive parallel sequencing, although the high costs of this method currently prevent large scale studies. In order to perform powerful and cost-effective studies, several strategies are discussed that combine massive parallel sequencing with other existing methods and techniques. Furthermore, the limitations of the different strategies are addressed, as well as the improvements needed in the near future to identify causative variants underlying a variety of traits in, but not limited to, the chicken.

Samenvatting

Wereldwijd wordt meer dan een kwart van het vlees en vrijwel alle eieren geproduceerd door kippen. Om in toekomstige voedsel behoeften te voorzien, het dierenwelzijn te verbeteren en te voldoen aan overige eisen van de consument, is het van belang om voor ziekten en productie eigenschappen de onderliggende genetische basis in kaart te brengen. Het doel van het onderzoek beschreven in dit proefschrift is het detecteren van genetische varianten die invloed hebben op, of de oorzaak zijn van, verschillende ziekten of productie eigenschappen in de kip. Om dit doel te bereiken hebben wij de bruikbaarheid van verschillende (nieuwe) moleculaire technieken onderzocht.

In **hoofdstuk 1** wordt een algemene introductie gegeven over de historie van de kip, met een nadrukkelijke focus op domesticatie en het opzetten van populaties die gespecialiseerd zijn in vlees en ei productie. Verder bevat dit hoofdstuk een overzicht van verschillende moleculaire technieken en methoden die gebruikt zijn, of kunnen worden, om genetische variatie te detecteren.

Koppelingskaarten zijn essentieel voor 'linkage mapping' studies, belangrijk om recombinatie binnen het genoom te bestuderen, en bij het reconstrueren van genoom sequenties (bv. een referentie genoom). In **hoofdstuk 2**, beschrijven wij een nieuwe, hoge resolutie, koppelingskaart van het kippen genoom gebaseerd op 1619 individuen van twee verschillende populaties. De twee populaties bestaan uit een vleeskippenlijn en een experimentele kruising tussen twee verschillende vleeskippenlijnen. De hoge resolutie van de koppelingskaart heeft er toe geleid dat wij recombinatie hotspots (plekken in het genoom waar veel recombinatie plaats vindt) in het genoom hebben kunnen opsporen. Bovendien hebben we voor 360 genetische merkers, waarvan de positie in het genoom onbekend was, een positie bepaalt in het genoom. Deze nieuwe gepositioneerde merkers kunnen gebruikt worden om het referentie genoom van de kip te verbeteren. De uiteindelijke koppelingskaart bestaat uit 13.340 SNP merkers en vertegenwoordigt 31 verschillende chromosomen. De totale lengte van de koppelingskaart voor de twee populatie tezamen is gemiddeld 3054 centiMorgans. De vergelijking tussen de koppelingskaarten van de twee afzonderlijke populaties toonde verschillende regio's op verscheidene chromosomen waarin de mate van recombinatie significant verschilde tussen de twee populaties. Deze regio's waren meestal gelegen op het telomerische uiteinde van de p-arm van een chromosoom. Op basis van de sekse specifieke koppelingskaart van beide populaties is gebleken dat de vrouwelijke recombinatie frequentie in de experimentele kruising afwijkend was.

In **hoofdstuk 3** beschrijven wij de moleculaire karakterisering van het 'late bevedering' gen, een enkelvoudig kenmerk in de kip dat leidt tot een vertraagde groei van veren (voornamelijk slagpennen) in eendagskuikens. Omdat dieren zonder het late bevedering gen een normale groei van de slagpennen hebben ('snelle beverding'), is er een duidelijk onderscheid te maken tussen kuikens die de verschillende gen varianten dragen. Omdat het gen op het sekschromosoom Z ligt, is het mogelijk om kruisingen te maken waarbij, bijvoorbeeld, alle kuikens met een langzame bevedering mannetjes zijn, terwijl alle kuikens met een snelle beverding vrouwtjes zijn. Doordat het 'late beverding' gen gebruikt kan worden voor sekse typering bij eendagskuikens wordt het gen veel gebruikt in de fokkerij. Het gen heeft echter ook een nadeel door de aanwezigheid van een dichtbij gelegen retroviraal virus. Dit retrovirus wordt vrijwel altijd samen gevonden met het gen voor late beverding. Het retrovirus heeft echter een nadelig effect op ei productie en veroorzaakt een verhoogd risico op virale infecties. Het identificeren van het 'late beverding' gen zal er toe kunnen leiden om dieren te vinden die wel het gewenste 'late bevedering' gen dragen, maar niet het ongewenste retrovirus. In dit hoofdstuk hebben wij met behulp van kwantitatieve PCR een duplicatie van ongeveer 180.000 basenparen ontdekt dat de late bevedering veroorzaakt. Deze duplicatie resulteert in een (gedeeltelijke) duplicatie van twee genen die coderen voor de prolactine receptor (*PRLR*) en sperm flagellar protein 2 (*SPEF2*). Op basis van de sequentie analyses hebben wij een moleculaire test ontwikkeld waarmee getest kan worden of een dier een drager is van deze duplicatie. Samen met bestaande testen is het nu mogelijk om dieren te identificeren die wel het gewenste 'late bevedering' gen dragen, maar niet het ongewenste retrovirus.

Door de recente ontwikkelingen in nieuwe (2^{e} generatie) sequentie technieken is het betaalbaar geworden om vrijwel alle genetische variatie binnen een genoom te detecteren binnen één enkel experiment. Door de 2^{e} generatie sequentie techniek is het gemakkelijker geworden om genetische varianten op te sporen die geassocieerd zijn met meervoudige kenmerken (een kenmerk of ziekte waarbij meerdere genen, elk met een kleine invloed op de ziekte, betrokken zijn). In **hoofdstuk 4** hebben wij de 2^{e} generatie sequentie techniek gecombineerd met een genoomwijde associatie studie om genetische varianten te detecteren die invloed hebben op Pulmonary Hypertension Syndrome (PHS). PHS, ook wel Ascites genoemd, is een ziekte bij de kip die beïnvloed wordt door meerdere genen. PHS is de afgelopen jaren in frequentie toegenomen, waarschijnlijk door de selectie op snelle groei en efficiënte voedsel conversie in de vleeskip. Momenteel is PHS één van de grootste oorzaken van diersterfte in vleeskippen, hetgeen niet alleen financiële gevolgen heeft voor de fokkerij bedrijven en boeren, maar ook leidt tot

een verminderd dierwelzijn. Met behulp van de genoomwijde associatie studie hebben wij 18 chromosomale regio's gevonden die mogelijk een invloed hebben op de ziekte. Om de onderliggende genetische oorzaak te vinden voor deze regio's hebben wij door middel van de 2e generatie sequentie techniek alle puntmutaties bepaald in het genoom van 12 individuen. Om zoveel mogelijk genetische varianten te detecteren hebben wij 6 individuen geselecteerd met veel zieke nakomelingen, en 6 individuen met veel gezonde nakomelingen. In 8 van de 18 chromosomale regio's hebben wij in totaal 10 genen opgespoord die genetische varianten bevatten die er toe leiden dat het eiwit (het eindproduct van een gen) niet meer goed kan functioneren. Selectie tegen deze varianten kan mogelijk tot een afname van PHS leiden.

Regio's in het genoom waar selectie heeft plaatsgevonden bevatten naar alle waarschijnlijkheid genen die van invloed zijn op productie kenmerken of ziekten. Het opsporen van deze genen kan bijdragen tot het verbeteren van productie kenmerken en het verhogen van resistentie tegen bepaalde ziekten. In **hoofdstuk 5** hebben wij de techniek 'hitch-hiking mapping' toegepast om in een groot aantal kippenrassen regio's te detecteren die onder selectie hebben gestaan. Wij hebben hiervoor niet alleen gekeken naar commerciële rassen (vlees- en legkippen), maar ook naar niet commerciële rassen uit Nederland en China. Bij 67 verschillende populaties binnen deze verschillende rassen hebben wij met behulp van een DNA SNP-chip (57.636 merkers) gekeken naar chromosomale regio's met significante aanwijzingen voor selectie in het verleden. Ons onderzoek demonstreert de toegevoegde waarde van het gebruik van meerdere populaties per ras, en van verschillende rassen met dezelfde fokdoelen, om zodoende stochastische effecten te verminderen. Binnen de chromosomale regio's die mogelijk onder selectie hebben gestaan, hebben wij verschillende genen gevonden die ook in de toekomst mogelijk van belang zijn voor het verbeteren van productie kenmerken en ziekte resistentie.

In de algemene discussie in **hoofdstuk 6** beschrijf ik de belangrijkste resultaten van dit proefschrift. In dit hoofdstuk worden aanbevelingen gedaan voor de beste strategieën om genetische varianten te detecteren voor zowel enkelvoudige als meervoudige kenmerken. Vrijwel alle strategieën kunnen profiteren van de 2e generatie sequentie technieken, al zal het uitvoeren van grootschalige experimenten momenteel nog gelimiteerd zijn door de hoge kosten van deze techniek. Bovendien bespreek ik in dit hoofdstuk de beperkingen van alle genoemde strategieën, en bediscussieer ik de verbeteringen die nodig zijn voor de toekomstige detectie van genetische varianten bij de kip en andere landbouwhuisdieren.

Dankwoord

Dankwoord

Het is dat de inhoud van de eerste vier pagina's van mijn proefschrift verplicht zijn, anders had dit dankwoord daar gestaan. Het dankwoord is natuurlijk het belangrijkste gedeelte van het proefschrift, want zonder hulp is het bijna onmogelijk om een proefschrift succesvol af te ronden. Zo ook voor mij. Zonder de goede werksfeer in onze groep, en natuurlijk hulp en gezelligheid van vrienden en familie, zou mijn promotie traject een stuk minder leuk en succesvol zijn geweest.

Tijdens het werk waren voor mij de koffiepauzes (ongeacht de kwaliteit van de koffie) een bron van ontspanning. Vooral zeer interessant waren de fantastische discussies over de tuinklauw goud, het smakelijk oerbrood, het wel of niet meedoen aan "Ter land, ter zee en in de lucht", hot-swappen, kerstmarken, kippen, de TRA, de portal, de koffiekwaliteit, de kleur van het nieuwe tapijt, en natuurlijk voetbal (helaas ben ik wel de enige binnen de vakgroep met échte voetbalkennis). Ontspannende evenementen zoals de bieravonden, de we-day en personeelsuitjes hebben natuurlijk ook bijgedragen aan de geweldige sfeer binnen onze groep (ok, het was niet altijd even succesvol, en om maar te zwijgen over de lichamelijk gevolgen, of het moedwillig vermoorden van clubmascottes). Waarschijnlijk zal ik in de tekst hieronder niet iedereen noemen die bij heeft gedragen aan de werksfeer. Maar mocht je je aangesproken voelen door bovenstaande tekst: DANK!

Natuurlijk zijn er ook vele personen die direct mee hebben geholpen aan mijn proefschrift. Allereerst wil ik mijn directe begeleiders Martien en Richard heel erg bedanken voor de kans die ik heb gekregen om na mijn afstudeervak bij jullie aan de slag kon gaan als AIO. Bedankt voor alle discussies, commentaren en vooral de motiverende woorden als het even tegen zat. Johan, ook al ben je niet direct betrokken bij mijn gedeelte van het project, bedankt voor alle hulp.
Natuurlijk was het ook niet gelukt zonder de rest van de 'Ascites groep'. Pieter, we hebben wat afgezien in de stallen, maar het is ons toch gelukt met de phenotyperingen! Ook dank voor het uitvoeren van de genotyperingen in Utrecht. Het is altijd knap om een protocol van 3 dagen terug te kunnen brengen naar 2 dagen! Veel plezier en succes bij Hendrix Genetics (maar dat zit wel goed geloof ik)! Ane, as a fellow-PhD-student-in-crime for the Ascites project you helped me tremendously with the quantitative side of the project, and off course with the 'inspiring' work of heart cutting. Good luck with finishing your thesis! Henk, heel erg bedankt voor je hulp. Gelukkig was jij rond een uur of 6 altijd nog aanwezig om

mijn vragen over statistiek te beantwoorden. Ook heb jij natuurlijk veel bijgedragen aan het 'kwantitatieve gedeelte' van het project.
Ook wil ik hierbij Addie en Gerard (Hendrix Genetics Research, Technology & Services B.V.) en Gosse (Cobb-Vantress) bedanken voor hun bijdrage aan het project. Addie, jouw kennis en inzet voor het project zijn van zeer grote waarde geweest. Bedankt voor alle snelle antwoorden als ik weer een vraag had over de historie van een populatie.
Het meeste werk en eigenlijk het belangrijkste deel van het project zijn de phenotyperingen in de 'stallen' geweest. Natuurlijk waren deze niet goed verlopen zonder de inbreng van vele personen. Johanna (ongelooflijk hoeveel werk jij hebt gedaan), Ger (onze fantastische weger), Ginny, Jan-Willem, Gerard, Birgitte, Piet, Alex, Monique. De basis van een goed experiment begint bij de basis van het experiment. Dankzij jullie nauwkeurige werk hebben wij een fantastische dataset verkregen. Heel erg veel dank daarvoor! Speciale dank gaat ook naar Annie en Theo. Door jullie gastvrijheid kwam ik graag om de twee dagen naar 'mijn kippetjes' kijken. Annie, bedank voor de gezellige praat, en de lekkere koffie, koekjes, en soep (kippensoep natuurlijk!). Ik hoop dat de navigatie het nog doet in je auto. Theo, met jouw (praktijk) kennis, nuchterheid, en gedrevenheid heb je ons erg geholpen.

Naast het werk in de stallen is het voor een 'molbi' natuurlijk ook belangrijk de juiste hulp te hebben in het lab. Zonder de 'lab mensen' was alles zeker een stuk minder leuk geweest. Bert (super dat je naast me staat op het podium!), Tineke (het heeft echt geen zin om te schelden tegen een machine!) hartelijk dank voor het uitvoeren van de 'DNA chips' in Utrecht en alle andere dingen die ik niet kan noemen omdat ik dan over een week nog niet klaar ben met dit dankwoord. Natuurlijk ook dank voor de andere mensen op het lab: Rosilde (ja voor mij val je ook onder het lab!), Sylvia, Kaveh, en Jan-Willem.

Verder zijn er natuurlijk nog de andere 'ondersteunende' mensen die ik wil bedanken. Annemieke, mede dankzij jouw begeleiding is mijn afstudeervak goed verlopen, wat er toe heeft geleid dat ik als AIO kon beginnen binnen de leerstoelgroep. Hendrik-Jan, vooral het afgelopen jaar heb ik je lastig gevallen met vragen over de nieuwe servers, programmeren, NGS, maar ook 'normale' (vooral evolutionair georiënteerde) vragen. Ook al had je vaak (eigenlijk) geen tijd, een simpele vraag leidde al snel tot een levendige discussie van een uur. Wie weet eten we samen nog een keer een Atlasburger in San Diego! Hinri, jij bent misschien wel de grootste inspirator geweest voor de onderwerpen tijdens de koffiepauze (laten we het maar niet hebben over het beestachtig en moedwillig molesteren van

collega's). Ook bedankt voor je eindeloze hulp betreffende de Bioinformatix, en het ondersteunen van alle ins en outs van Microsoft Windows (of was het toch dat andere systeem met die pinguïn?). Aniek, bedankt voor de hulp van de R-plotjes in mijn proefschrift! Katrijn, bedankt voor de ontspannende (althans voor iedereen die niet aan het koken was) en bovendien heerlijke etentjes! Ik ben blij dat je ook naast me durft te staan in de aula. Ilse, bedankt voor de nuttige tips gegeven over de cover foto. Ik had ook niet anders verwacht van de medeoprichter van x-pressions.nl (Kijk uit voor die andere oprichter, die wil nog wel eens wat slopen). Amélie, thanks for helping me on chapter 3. Good luck with your PhD!

Het leven van een promovendus bestaat (over het algemeen) natuurlijk niet alleen uit werk. Veel dank gaat uit naar de 'spoeften' uit Nijmegen. Chantal, Femke, Jeroen, Petra, Rob, het is raar hoeveel impact een jaar samenwonen in een zusterflat (6 m^2 per kamer!) kan hebben op mijn leven. Samen met jullie heb ik één van de gezelligste jaren uit mijn leven gehad. Ik vind het nog steeds fantastisch dat we elk jaar weer ons traditionele weekendje CP hebben (bovendien met een toenemende populatie!).

Ook wil ik hierbij mijn 'WaJo vrienden' (Aage, Derk-Jan, Edo, Jasper, Lianne, Marjolein, Martijn, Paul, Rene en Rik) bedanken. Bedankt voor de gezamenlijke etentjes op maandag (pasta, bonenschotel óf nasi), sporten (naja…zoiets), fifa-avonden (ik ben in potentie nog steeds de beste!), discussies over voetbal (gelukkig heeft J wél voetbalkennis), en natuurlijk het stappen!

Wim, Bep, Esther en Thij, bedankt voor alle gezelligheid!

Pa en Ma, zonder jullie was dit proefschrift nooit tot stand gekomen! En dan bedoel ik niet alleen dat ik zonder jullie niet geboren zou zijn. Door middel van jullie opvoeding, hulp, motivatie en inzet ben ik geworden wie ik ben. Dankzij jullie heb ik de kans gekregen om te studeren, en wederom te studeren, met dit proefschrift als uiteindelijk gevolg. Bedankt voor alles!
Ingmar, bedankt voor alle hulp om mijn bolides rijdend te houden, en voor het gezelschap tijdens de fantastische wedstrijden van Vitesse!

Als laatste wil ik mijn lieve schat Marion bedanken. Enkele weken voordat ik aan dit project begon, leerde ik jou kennen. Bedankt voor alle steun, gezelligheid en leuke momenten samen. In tegenstelling tot mijn AIO project zal ons gezamenlijk project gelukkig nog lang niet eindigen!

About the author

Curriculum vitae

Martin Gerhard Elferink was born on the 11th of September, 1980 in Losser (the Netherlands) and was raised in Rhenen. In 1998 he obtained his high school diploma (HAVO) at the O.S.G. Pantarijn in Wageningen. In 1998, Martin started his study on molecular biology at the H.L.O. of the Hogeschool van Utrecht. He obtained his bachelor's degree in 2002 after completing his thesis at the cytogenetic division of the department of Human Genetics, Radboud University, Nijmegen. After obtaining his bachelor's degree he worked as a research technician at the department of Human Genetics from the Radboud University (Nijmegen) on the development of a 32k BAC array used for array comparative genomic hybridization. In 2004, Martin became a research technician at the department of Plasma Proteins, Sanquin B.V. (Amsterdam), where he worked on the isolation and stabilization of coagulation Factor IX. In 2004, he started his Master in Bioinformatics at the Wageningen University where he obtained his degree in 2006 after completing his thesis at the Animal Breeding and Genomics Centre, Wageningen University. After obtaining his Master's degree Martin started his PhD project "The characterisation of genes involved in pulmonary hypertension syndrome in chicken" at the Animal Breeding and Genomics Centre, a project that was funded by the Dutch Technology Foundation (STW). This project resulted in this thesis. Currently, Martin is continuing his scientific career as a Postdoctoral research associate at the Animal Breeding and Genomics Centre, Wageningen University.

Curriculum vitae

Martin Gerhard Elferink werd geboren op 11 september 1980 in Losser (Nederland) en groeide op in Rhenen. In 1998 behaalde hij zijn HAVO diploma aan het O.S.G. Pantarijn te Wageningen. Datzelfde jaar begon hij aan de studie Moleculaire Biologie aan de H.L.O. van de Hogeschool van Utrecht. Deze studie werd succesvol afgerond in 2002, na een stage bij de sectie Cytogenetica van de afdeling Human Genetics van de Radboud Universiteit te Nijmegen. Na het behalen van zijn H.B.O. diploma is Martin als research analist blijven werken bij de afdeling Human Genetics waarin hij werkte aan het ontwikkelen van een 32k BAC array voor array comparative genomic hybridization. In 2004 begon Martin als research analist bij de afdeling Plasma eiwitten van Sanquin B.V., waarin hij werkte aan de isolatie en stabilisatie van stollingsfactor IX. In 2004 begon hij zijn Master opleiding in bioinformatica aan de Universiteit van Wageningen. In 2004 werd deze studie succesvol afgerond na het uitvoeren van een afstudeervak bij de afdeling Animal Breeding and Genomics Centre, Wageningen University. Na het behalen van zijn Masters diploma is hij als AIO begonnen binnen dezelfde groep aan het project “The characterisation of genes involved in pulmonary hypertension syndrome in chicken” dat gefinancierd was door STW. De resultaten van dit project zijn beschreven in dit proefschrift. Momenteel zet Martin zijn wetenschappelijke carrière voort als Postdoctoral research associate bij de afdeling Animal Breeding and Genomics Centre aan de universiteit van Wageningen.

List of publications

Peer-reviewed publications

M.G Elferink, A.A.A. Vallée, A.P. Jungerius, R.P.M.A. Crooijmans, and M.A.M. Groenen. Partial duplication of the *PRLR* and *SPEF2* genes at the late feathering locus in chicken. *BMC Genomics*. 2008; Aug 20;9:391.

M.G. Elferink, P. van As, T. Veenendaal, R.P.M.A. Crooijmans, M.A.M. Groenen. Regional Differences in Recombination Hotspots between Two Chicken Populations. *BMC Genetics*. 2010; Feb 8;11:11.

P. van As, **M.G. Elferink**, A.M. Closter, A. L. J. Vereijken, H. Bovenhuis, R.P.M.A. Crooijmans, E. Decuypere and M.A.M. Groenen. The Use Of Blood Gas Parameters To Predict Ascites Susceptibility In Juvenile Broilers. *Poultry Science* 89 :1684–1691.

Publications in review or preparation

M.G. Elferink, H-J. Megens, A. L. J. Vereijken, X. Hu, R.P.M.A. Crooijmans, M.A.M. Groenen. Signatures of selection in the genome of commercial and non-commercial chicken breeds. *Manuscript submitted*

M.G. Elferink, A.M. Closter, P. van As , D. Nikolic, H-J Megens, H. Bovenhuis, R.P.M.A. Crooijmans, M.A.M. Groenen. Massive parallel sequencing of 12 genomes identifies protein affecting variants within QTL regions associated with the pulmonary hypertension syndrome in chicken. *Manuscript in preparation*

A. M. Closter, **M. G. Elferink**, P. van As, R. P. M. A. Crooijmans, A. L. J. Vereijken, A. M. van Arendonk, M. A. M. Groenen, H. Bovenhuis. Genetic correlation between heart ratio and body weight as a function of ascites incidence. *Manuscript in preparation*

Conference proceedings

M.G. Elferink, A.M. Closter, P. van As, H. Bovenhuis, A. Vereijken, G. Veninga, R.P.M.A. Crooijmans, M.A.M. Groenen. The characterisation of genes involved in the Pulmonary Hypertension Syndrome in chicken. *Book of abstracts* of the Annual ZonMw Genetica Retraite, Kerkrade, the Netherlands (2007).

M.G. Elferink, A.A.A. Vallée, A. Jungerius, R.P.M.A. Crooijmans, M.A.M. Groenen. Partial duplication of *PRLR* and *KPL2* at the late feathering locus in chicken. *Book of abstracts* of the International Society for Animal Genetics, Amsterdam, the Netherlands (2008).

M.G. Elferink, A.M. Closter, H. Bovenhuis, P. van As, A. Vereijken, R.P.M.A. Crooijmans, M.A.M. Groenen. Genome-wide association study of pulmonary hypertension syndrome in chicken. *Book of abstracts* of the Annual ZonMw Genetica Retraite, Kerkrade, the Netherlands (2010).

M.G. Elferink, A.M. Closter, P. van As , H. Bovenhuis, A. Vereijken, G. Veninga, R.P.M.A. Crooijmans, M.A.M. Groenen. Genome-wide association study of pulmonary hypertension syndrome in chicken. *Book of abstracts* of the Plant & Animal Genome XVIII, San Diego, USA. (2010).

Training and supervision plan

The Basic Package (3 ECTS)

WIAS Introduction Course	2007
Course on philosophy of science and/or ethics	2007

Scientific Exposure (14.5 ECTS) *Conferences, seminars, and workshops*

Basic Real Time PCR training	2006
WIAS Biodiversity seminar	2006
EADGENE and SABRE "Genomics for Animal Health"	2007
International Society for Animal Genetics	2008
F&G connection dagen	2008
GRC Quantitative genetics and genomics	2009
WIAS science day	2009
Plant & Animal Genome XVIII	2010
Annual ZonMw Genetica Retraite	2007,2008,2010

In-Depth Studies (8.5 ECTS)

Understanding Genotype Enviroment Interactions	2007
QTL Mapping, MAS, and Genomic Selection	2008
Advanced sequencing technologies & applications, CSHL	2009

Professional Skills Support Courses (3.6 ECTS)

Writing winning grant proposals	2006
Introduction to R for statistical analysis	2008
Techniques for writing and presenting a scientific paper	2009
Project and Time Management	2010

Research Skills Training (2 ECTS)

Research proposal "Characterisation of the early and late feathering locus'	2006

Didactic Skills Training (14.4 ECTS)

Teacher assistant Genomics course (ABG-30306)	2007-2010
ABG course, supervision practical's	2008
RMC course reviewer	2009
Supervising 3 MSc students	2006,2008,2010
Supervising 1 MLO student (internship)	2009

Total credits (ECTS): 46

Colophon

The research described in this thesis is financially supported by the Dutch Technology Foundation (STW). Project: “The characterisation of genes involved in pulmonary hypertension syndrome in chicken”. Project number 07106.

Printed by: GVO drukkers & vormgevers B.V. | Ponsen & Looijen, Ede, the Netherlands