Genome-wide Gene Expression Surveys and a Transcriptome Map in Chicken

Thesis committee

Thesis supervisors

Prof. dr. M.A.M. Groenen
Personal chair at the Animal Breeding and Genetics Groups
Wageningen University

Prof. dr. M.A. Smits
Personal chair at the Animal Breeding and Genetics Group
Animal Science Group (ASG), Lelystad
Wageningen University and Research Centre

Thesis co-supervisor

Dr. R.P.M.A. Crooijmans
Assistant professor, Animal Breeding and Genetics Group
Wageningen University

Other members

Prof. dr. M.K. Richardson (Leiden University)
Prof. dr. N. Li (China Agricultural University, Beijing, China)
Prof. dr. M.R. Muller (Wageningen University)
Prof. dr. J. Keijer (Wageningen University)

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Genome-wide Gene Expression Surveys and a Transcriptome Map in Chicken

Haisheng Nie

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 8 March 2010
at 4 p.m. in de Aula.

Haisheng Nie (2010)

Genome-wide Gene Expression Surveys and a Transcriptome Map in Chicken

Ph.D. Thesis, Wageningen University, The Netherlands

With summaries in English and Dutch

ISBN: 978-90-8585-622-1

This thesis is dedicated to my parents for their limitless love and support

Contents

Chapter 1

General introduction

1.1 Chicken – an important model organism in biological research

The chicken (*Gallus gallus*) is an important model organism in genetics, developmental biology, immunology, and evolutionary research. Moreover, besides being an important model organism the chicken is also a very important agricultural species and an important source of food (eggs and meat).

The chicken started to being used as a model organism in genetics more than one hundred years ago and terms such as alleles [1], genetic linkage [2] and epistasis [3] are based on work on chicken morphological traits [4]. Over the years chicken genetics has mainly focused on practical problems of meat and egg production and on the analysis of disease resistance. A recent literature search of the PubMed [5] database, using the key words, "chicken" and "genetics", returned more than 14,000 records (August 2009), clearly indicating the intensive use of chicken in the field of genetics. In addition to genetic studies related to more practical agriculturally related aspects, the chicken for many decades has intensively been used to study embryonic development. This extensive use of the chicken as one of the primary models for developmental biology is due to the easy access of the embryo because development occurs in ovo rather than in utero, which allows easy manipulation of the incubated eggs and the developing embryo. Chicken has also been used intensively in immunological research. The chicken immune system provided the first distinction between two different types of immune cells, T-cells and B-cells. The B-cell itself was named based on the chicken bursa of Fabricius [6].

1.2 The chicken genome and chicken genomics researches

The chicken is one of the non-mammalian vertebrate model organism and it shares the last common ancestor with mammals about 310 million years ago [7]. Because of its importance as a model organism for developmental biology and agriculture and because of its phylogenic distance from mammals the chicken (*Gallus gallus*) genome was sequenced in 2004 [8]. Because of its strategic evolutionary position in the tree of life between mammals and fish, chicken is an important anchor species in the phylogenetic study of genome evolution. Sequencing the chicken genome also helps to improve our understanding of the functioning of mammalian genomes including human, through comparative genomics. The chicken mammalian comparison has a high signal-to-noise ratio resulting in a high specificity for the identification of regions under selection (conserved during evolution). A comparison between human and chicken showed that 75% of the coding regions and 30%-40% of regulatory elements are conserved when examining known functional sequences between mammals and chicken [9].

The availability of the draft chicken genome sequence [8] provided many possibilities to in detail study a variety of genomic changes during evolution using a comparison between chicken and mammals. For example, compared to mammals, the use of a Z/W sex determination system is a special aspect of the avian genome, where the female is the heterogametic sex (ZW) and the male is the homogametic (ZZ) sex. A comparison of the genomic sequences of platypus, chicken, and human showed that sex chromosomes evolved separately in birds and mammals [10]. As mentioned before, the evolutionary position of chicken compared to mammals results in a high specificity for the detection of functional elements in vertebrate genomes [11-13]. A clear example is provided by the ultra conserved sequences often co-localizing with developmental genes (including genes linked to disorders that cause limb loss or deformity) [8]. The draft chicken genome sequence also provided several interesting biological observations. For example, the observed number of olfactory receptors in the chicken genome challenged the fact that the chicken has been thought to have a poor sense of smell. The number of genes in the chicken genome coding for olfactory receptors is similar as is found in the human genome which suggest that the chicken has a sense of smell more or less similar as human [8].

Another interesting feature of the chicken genome is the great variation in size of the different chromosomes. This karyotype consisting of both large (macro) as well as small (micro) chromosomes is very characteristic for most avian genomes [14]. Microchromosomes are also found in some primitive amphibians [15, 16] and most reptiles [17]. Most avian karyotypes are composed of about 40 pairs of chromosomes. Some notable exceptions are the stone curlew and kingfisher, with 20 and 66 pairs of chromosomes, respectively [18]. Interestingly, microchromosomes exhibit higher gene density, smaller gene size, and higher recombination rates compared to macrochromosomes [8, 19].

The important evolutionary position of chicken relative to other mammals makes chicken an interesting model in the current genomics research to address several basic, yet important genomic questions, such as the evolution of genome size [9]. In recent years, several genomic resources were developed for chicken, such as a high-density SNP-based linkage map [19], a 50K SNP i-sellect panel (Illumina), genome-wide expression microarrays (ARK-Genomics G. gallus 20K oligo array, chicken 44k Agilent array [20]) and CNV (Copy Number Variation) arrays (Nimblegen 385k tiling path array and Agilent 244k chicken array). These resources and high-throughput platforms provide the necessary tools to further investigate the chicken genome in more detail.

Since the first draft of the chicken genome sequence (WASHUC1) released in 2004, a newer assembly is available since 2006 (WASHUC2). However, in both builds several microchromomes are still poorly covered and the ten smallest microchromosomes are still completely missing. Recently the same red jungle fowl (UCD001) was re-sequenced at Washington University using 454 sequencing technology (Roche). This next-generation sequencing technology avoids the bacterial cloning steps of required using traditional Sanger sequencing and is expected to provide better coverage of the chicken genome, especially the microchromosomes. This new assembly is available at http://genome.wustl.edu/genomes/view/gallus_gallus/ and has an overall sequence depth of 19x (6.6 x for WASHUC2 and 12x Roche 454 sequences). This new assembly, together with a collection of available genomic tools will further strengthen the usefulness of chicken as a popular model in the future genomics research by providing more and better genomic data for chicken biology.

1.3 Transcriptomics research in chicken

The sequence of the chicken genome has provided new possibilities to study the function of the individual genes and gene networks in chicken and to gain insight in their specific roles in chicken physiology. An important challenge in the post-sequence era of chicken biology is determining the functional role of known genes. Currently, the function of many of the chicken genes has been predicted based on the sequence homology to genes of know function in other species. However, because of the differences in physiology among different species it is essential to improve this and to obtain additional functional data in the chicken itself, for example, more detailed information about the expression of these genes in different tissues and under different conditions.

Before 2003, only a few papers on gene expression profiling using microarrays were published in chicken [21]. The first picture of global gene expression in the immune system in chicken was provided by lymphoid cDNA microarrays [22]. Subsequently, several tissue-specific cDNA microarrays were developed and used for transcription profiling in the liver, intestine and bursa of Fabricius [23-25]. The first high-density (13K) multi-tissue chicken cDNA array to be developed [26], was based on ESTs/cDNA clones representing 24 different adult or embryonic tissues. The coverage of cDNA microarray platforms increased very fast during the following years and, at the same time, the quality of array manufacturing also improved. The availability of a draft chicken genome sequence in 2004, made it possible to manufacture whole genome oligo-arrays to study genome-wide gene expression in chicken. The Chicken Genome GeneChip, containing probes for 33,457 chicken and viral pathogen transcripts, is commercially available from Affymetrics (http://www.affymetrix.com) (GEO [27] accession: GPL3213) and this array was the first genome-wide gene expression chip on the market. Microarrays consisting of long oligonucleotides (70-mer) were developed by a number of different groups including the Roslin Institute (ARK-Genomics G. gallus 20K; GEO accession: GPL5480, GPL8862), the University of Arizona (*Gallus gallus* 20.7K Oligo Array; GEO accession: GPL6049) and the University of Missouri (*Gallus gallus* 21k; GEO accession: GPL5618). Two of these microarrays are available from the University of Arizona (http://www.grl.steelecenter.arizona.edu/) and ARK Genomics (http://www.ark-genomics.org/), Recently, a Chicken 44K custom Agilent microarray (GEO accession: GPL4993, GPL7399, GPL8764) was developed which currently is available from Agilent (http://www.agilent.com/).

By the time, the experiments for this thesis were carried out, the Agilent platform was not available yet, and the cost of Chicken 20k oligo-arrays was much lower than the Affymetrix chips. This made it a preferable platform to use within our project, because of the relatively large number of samples to survey. Nowadays, these commercial long-oligo arrays and the chicken genome array (GeneChip) are increasingly replacing custom microarrays (for example, tissue specific cDNA microarrays) because of the higher standardization and higher quality of these platforms. The current publicly available transcriptomic data in chicken using genome-wide oligo array is shown in table 1.

Table 1. An overview of genome-wide expression studies in chicken in NCBI GEO database.

Platform Accession ID	Description	Platform Name	No. of dataset	No. of arrays
GPL3213	GeneChip	Affymetrix Chicken Genome Array	28	404
GPL4993	Agilent 44K oligo set	Chicken 44K custom Agilent microarray	3	60
GPL7399	---	Agilent custom 44K chicken array	1	20
GPL5480	Roslin/Ark 20K oligo set	ARK-Genomics G. gallus 20K v1.0	2	64
GPL5618	---	Missouri Gallus gallus 21k	1	9
GPL6049	---	Arizona Gallus gallus 20.7K Oligo Array v1.0	2	120
GPL8199	---	ChickenOligo 20.6K 70-mer microarray v2	2	16

The number of genome-wide transcription profiling experiments in chicken has increased dramatically in recent years because of the availability of the microarray platforms described above. The availability of these new platforms of improved quality and higher probe coverage, for example the Agilent 44K chicken array, we expect that the number of genome-wide transcription profiling studies using these platforms will increase in the near future as well, studies that use direct sequence-based technology to study gene expression.

The majority of previous microarrays studies described in GEO or other publications [28-34] were designed to monitor changes of gene expression between different conditions, treatments, or time points in a single tissue. The identified candidate genes were subsequently used to try to interpret the underlying biological processes by looking at the functions of these genes. There are many candidate genes identified that lack any functional annotation in the current chicken genome build, hampering the interpretation of the results obtained within these microarray experiments.

This limitation of the current microarray analysis motivated the generation of transcriptional profiling across a number of tissues in project described in this thesis. The global expression pattern of the genes under normal conditions among tissues can be used as a reference baseline for expression studies aimed to study specific diseases in chicken. It provides information about the distribution of the gene transcription profile across a range of tissues under normal conditions and this will facilitate the inference of possible biological functions of un-annotated genes in chicken. Genome-wide gene expression information in chicken can also be used to shed light on other aspects of vertebrate genome and transcriptome evolution. For example, it has been reported in human [35, 36], that housekeeping genes have relatively shorter introns, untranslated regions and coding sequences, suggesting a selection for compactness. With genome-wide gene expression data in chicken across a number of tissues, we can identify genes with “housekeeping functions” and test whether the compactness of housekeeping gene found in human is also true in chicken. Furthermore, evolutionary changes in gene expression account for most phenotypic differences between different species. Global gene expression patterns were reported to be conserved between human and apes [37] as well as between human and mouse [38]. The results of these studies suggested that the gene expression within mammals is under evolutionary constraint. Comparing gene expression of birds and mammals would help to further understand the gene expression conservation in vertebrates during evolution.

1.4 Gene transcription regulation

Regulation of transcription is known to be regulated at a number of different levels, i.e. at the individual gene level, at the level of gene clusters, and at a more global regional genomic level. The first level of regulation is on individual genes. This common model for eukaryotic gene transcription involves the binding of several transcription factors (TFs) to promoter regions, resulting in activation of the individual genes. A good example is the well known TATA binding protein that regulates gene expression by binding to TATA box located in gene promoter regions [39]. The second level of gene regulation is on gene clusters. Most notably are the well-studied examples of a number of tightly co-regulated gene clusters, such as the globin, MHC and the Hox gene clusters [40-43]. For instance, the expression of MHC class II genes is tightly regulated at multiple levels of control by a series of cis-regulatory DNA elements interacting with transcription proteins or factors [44]. A third level of gene expression depends on the genomic locations of the genes [45-48]. This implies that genes located within the same region of the genome are co-regulated on a more global regional basis, beyond the level of functionally related gene clusters. In the human genome highly expressed genes appear to be clustered within specific chromosomal regions [49]. Further studies using specific insertions of GFP (green fluorescent protein) reporter gene constructs into these specific chromosomal regions showed an increased GFP expression of these inserted reporter genes as well [50]. Besides gene transcription, other characteristics such as gene density, GC content, nuclear position and recombination have been shown to exhibit domain-like features and are correlated with gene transcription activity in the eukaryotic genomes [19, 51]. The causative nature of inter-correlations of these features is still under investigation, but all these phenomena lead to the hypothesis that gene transcription, on top of the individual gene level regulation, is regulated in a domain-wide manner within vertebrate genomes, closely correlated with other structural characteristics in the genome. The observed location of gene-dense chromosome and chromosomal regionswith highly expressed genes towards the center of the nucleus and the location of gene-poor and weakly expressed chromosomes towards the nuclear envelope in human [52] and chicken cells [53] provided some further evidences of the existing correlation between gene transcription and other genomic features.

Furthermore, enhancers, silencers, locus control regions (LCR) and epigenetic regulators such as matrix attachment regions (MARs) are also known to be involved in gene transcription regulations. Metazoan LCRs, enhancers and silencers activate or repress transcription of linked genes at distal locations. Most enhancers are located tens of kilobases from the genes they regulate, and some have even been found at distances of up to a megabase from their gene target [54-56]. Furthermore, enhancer and silencers have been shown to have the potential to activate/repress a number of neighboring genes within a large chromosomal region [57]. Another type of regulatory element, the matrix attachment regions have been reported to serve not only as static organizers of nuclear and chromosomal structure but also as potentially dynamic DNA elements that exert important regulatory functions on the expression of individual genes [58]. All these known regulators may act, at all three levels of regulation described above, within a complex network acting on the target genes in the vertebrate genome to achieve accurate regulation of gene transcription.

In order to confirm the universal existence of the global region-wide levels of transcriptional regulation in vertebrate genomes, additional analyses are needed in additional species besides human and mouse. In this respect, the chicken is an important anchor species that can provide improved insights on the identification of the conservation of such region-wide levels of gene transcription in the different genomes. As described in section 1.3, chicken microchromosomes have specific features including higher gene density, higher recombination rate and shorter genes compared to the macrochromosomes. The availability of genome-wide gene expression resources in chicken will enable us to further investigate the mechanisms of transcriptional regulation of vertebrate genes.

1.5 Aim and outline of this thesis

The research described in this thesis was aiming to build a gene expression atlas for chicken by surveying genome-wide gene expression across a collection of adult and embryonic tissues and different staged whole embryos. The two genome-wide gene expression data sets are used as i) an expression baseline under normal conditions in chicken in contrast to specific treatments; ii) a references for comparative analysis of transcriptomics between different species iii) a resources to further study the regulation of gene transcription in eukaryotes. A transcriptome map for chicken was built using the expression data generated in this research and used to further study the mechanisms of gene regulation in vertebrate genomes.

The outline of this thesis: Chapter 2 provides an introduction to microarray data analysis and different normalization and analysis processes are discussed. Furthermore, limitations of the (chicken) microarray platform are discussed. Chapter 3 provides a general guideline for extracting biological information from microarray data with particular focus on species with less well-annotated genomes, like those for farm animals, using R/Bioconductor [59, 60] packages. The enrichment of gene annotations for functional information as well as for genomic locations, are studied. Biological pathways for differentially expressed (DE) genes under different combination of treatments are identified. Chapter 4 describes a gene expression survey in eight chicken adult tissues. Tissue-specific and housekeeping genes are identified among the tissues included in the survey. Functional enrichment analyses show that tissue-specific genes are enriched with GO terms corresponding to the physiological functions of the organs. Furthermore, housekeeping genes are found to be more compact comparing to tissue-specific genes and the expression of mouse-chicken-frog orthologous genes are found to be conserved. In chapter 5, a gene expression survey in whole chicken embryos from different developmental stages and embryonic tissues is described. This expression survey provides an atlas of gene expression in important embryonic stages and the major embryonic tissues in chicken. Stage- and tissue-specific genes are identified, and similar to chapter 4, housekeeping genes are found to be more compact. Differentially expressed genes between embryos from different developmental stages are identified and discussed in detail. In chapter 6, the chicken transcriptome map for the different chromosomes is presented, where highly expressed genes are found to be clustered together. This feature is highly correlated with other genomic features, such as for example gene density and GC content. This chapter describes a higher order level of transcriptional regulation in chicken, which seems to be conserved during evolution between chicken and human. Finally, in chapter 7 the results obtained in this thesis are discussed in a more general way. Some limitations of the current technological platform (microarray) are discussed and some perspectives are given for chicken transcriptomics using next-generation sequencing technology.

References

1. Bateson W and Saunders E: **Experiments in the physiology of heredity**. *Rep Evol Comm R Soc* 1902, **1**:1–160.
2. Sutton WS: **The chromosomes in heredity**. *Biol Bull* 1903, **4**:231–251.
3. Bateson W and Punnett RC: **The inheritance of the peculiar pigmentation of the Silky fowl**. *J Genet* 1911, **1**:185–203.
4. Burt DW: **Emergence of the chicken as a model organism: implications for agriculture and biology**. *Poult Sci* 2007, **86**(7):1460-1471.
5. **PubMed:** [http://www.ncbi.nlm.nih.gov/pubmed/]
6. Cooper MD, Raymond DA, Peterson RD, South MA, Good RA: **The functions of the thymus system and the bursa system in the chicken**. *J Exp Med* 1966, **123**(1):75-102.
7. Hedges SB: **The origin and evolution of model organisms**. *Nat Rev Genet* 2002, **3**(11):838-849.
8. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.
9. Hughes AL, Piontkivska H: **DNA repeat arrays in chicken and human genomes and the adaptive evolution of avian genome size**. *BMC Evol Biol* 2005, **5**(1):12.
10. Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, Alsop AE, Grutzner F, Deakin JE, Whittington CM, Schatzkamer K et al: **Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes**. *Genome Res* 2008, **18**(6):965-973.
11. Ovcharenko I, Loots GG, Giardine BM, Hou M, Ma J, Hardison RC, Stubbs L, Miller W:

Mulan: **multiple-sequence local alignment and visualization for studying function and evolution**. *Genome Res* 2005, **15**(1):184-194.

12. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L: **Evolution and functional classification of vertebrate gene deserts**. *Genome Res* 2005, **15**(1):137-145.
13. Gordon L, Yang S, Tran-Gyamfi M, Baggott D, Christensen M, Hamilton A, Crooijmans R, Groenen M, Lucas S, Ovcharenko I et al: **Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions**. *Genome Res* 2007, **17**(11):1603-1613.
14. Rodionov AV: **Micro vs. macro: structural-functional organization of avian micro- and macrochromosomes**. *Genetika* 1996, **32**(5):597-608.
15. Morescalchi A, Odierna G, Olmo E: **Karyological relationships between the Cyptobranchid salamanders**. *Specialia* 1977, **15**:1579.
16. Morescalchi A, Odierna G, Olmo E: **Karyology of the primitive salamanders, family Hynobiidae**. *Experientia* 1979, **35**:1434-1436.
17. Mengden GA, Stock AD: **Chromosomal evolution in Serpentes: a comparison of G and C chromosome banding patterns of some Colubrid and Boid genera**. *Chromosoma* 1980, **79**:53-64.
18. Burt DW: **Origin and evolution of avian microchromosomes**. *Cytogenet Genome Res* 2002, **96**(1-4):97-112.
19. Groenen MA, Wahlberg P, Foglio M, Cheng HH, Megens HJ, Crooijmans RP, Besnier F, Lathrop M, Muir WM, Wong GK et al: **A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate**. *Genome Res* 2009, **19**(3):510-519.
20. Li X, Chiang HI, Zhu J, Dowd SE, Zhou H: **Characterization of a newly developed chicken 44K Agilent microarray**. *BMC Genomics* 2008, **9**:60.
21. Cogburn LA, Porter TE, Duclos MJ, Simon J, Burgess SC, Zhu JJ, Cheng HH, Dodgson JB, Burnside J: **Functional genomics of the chicken--a model organism**. *Poultry science* 2007, **86**(10):2059-2094.
22. Neiman PE, Ruddell A, Jasoni C, Loring G, Thomas SJ, Brandvold KA, Lee R, Burnside J, Delrow J: **Analysis of gene expression during myc oncogene-induced lymphomagenesis in the bursa of Fabricius**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(11):6378-6383.
23. Cogburn LA, Wang X, Carre W, Rejto L, Porter TE, Aggrey SE, Simon J: **Systems-wide chicken DNA microarrays, gene expression profiling, and discovery of functional genes**. *Poultry science* 2003, **82**(6):939-951.

24. van Hemert S, Ebbelaar BH, Smits MA, Rebel JM: **Generation of EST and microarray resources for functional genomic studies on chicken intestinal health**. *Animal biotechnology* 2003, **14**(2):133-143.

25. Neiman PE, Grbic JJ, Polony TS, Kimmel R, Bowers SJ, Delrow J, Beemon KL: **Functional genomic analysis reveals distinct neoplastic phenotypes associated with c-myb mutation in the bursa of Fabricius**. *Oncogene* 2003, **22**(7):1073-1086.

26. Burnside J, Neiman P, Tang J, Basom R, Talbot R, Aronszajn M, Burt D, Delrow J: **Development of a cDNA array for chicken gene expression analysis**. *BMC genomics* 2005, **6**(1):13.

27. **National Center for Biotechnology Information (NCBI) Gene Expression Omnibus**: [http://www.ncbi.nlm.nih.gov/geo/]

28. van Hemert S, Hoekman AJ, Smits MA, Rebel JM: **Gene expression responses to a Salmonella infection in the chicken intestine differ between lines**. *Vet Immunol Immunopathol* 2006, **114**(3-4):247-258.

29. van Hemert S, Hoekman AJ, Smits MA, Rebel JM: **Immunological and gene expression responses to a Salmonella infection in the chicken intestine**. *Vet Res* 2007, **38**(1):51-63.

30. Kim DK, Hong YH, Park DW, Lamont SJ, Lillehoj HS: **Differential immune-related gene expression in two genetically disparate chicken lines during infection by Eimeria maxima**. *Dev Biol* (Basel) 2008, **132**:131-140.

31. Kim DK, Lillehoj HS, Hong YH, Park DW, Lamont SJ, Han JY, Lillehoj EP: **Immune-related gene expression in two B-complex disparate genetically inbred Fayoumi chicken lines following Eimeria maxima infection**. *Poult Sci* 2008, **87**(3):433-443.

32. Morgan RW, Sofer L, Anderson AS, Bernberg EL, Cui J, Burnside J: **Induction of host gene expression following infection of chicken embryo fibroblasts with oncogenic Marek's disease virus**. *J Virol* 2001, **75**(1):533-539.

33. Wang HB, Li H, Wang QG, Zhang XY, Wang SZ, Wang YX, Wang XP: **Profiling of chicken adipose tissue gene expression by genome array**. *BMC genomics* 2007, **8**:193.

34. Desert C, Duclos MJ, Blavy P, Lecerf F, Moreews F, Klopp C, Aubry M, Herault F, Le Roy P, Berri C et al: **Transcriptome profiling of the feeding-to-fasting transition in chicken liver**. *BMC genomics* 2008, **9**:611.

35. Eisenberg E, Levanon EY: **Human housekeeping genes are compact**. *Trends Genet* 2003, **19**(7):362-365.

36. Vinogradov AE: **Compactness of human housekeeping genes: selection for economy or genomic design?** *Trends Genet* 2004, **20**(5):248-253.

37. Khaitovich P, Enard W, Lachmann M, Paabo S: **Evolution of primate gene expression**. *Nature reviews* 2006, **7**(9):693-702.
38. Liao BY, Zhang J: **Evolutionary conservation of expression profiles between human and mouse orthologous genes**. *Molecular biology and evolution* 2006, **23**(3):530-540.
39. Lifton RP, Goldberg ML, Karp RW, Hogness DS: **The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications**. *Cold Spring Harb Symp Quant Biol* 1978, **42 Pt 2**:1047-1051.
40. Kielman MF, Smits R, Devi TS, Fodde R, Bernini LF: **Homology of a 130-kb region enclosing the alpha-globin gene cluster, the alpha-locus controlling region, and two non-globin genes in human and mouse**. *Mamm Genome* 1993, **4**(6):314-323.
41. The MHC sequencing consortium: **Complete sequence and gene map of a human major histocompatibility complex**. *Nature* 1999, **401**(6756):921-923.
42. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL et al: **Zebrafish hox clusters and vertebrate genome evolution**. *Science* 1998, **282**(5394):1711-1714.
43. Garcia-Fernandez J: **The genesis and evolution of homeobox gene clusters**. *Nat Rev Genet* 2005, **6**(12):881-892.
44. Glimcher LH, Kara CJ: Sequences and factors: **a guide to MHC class-II transcription**. *Annu Rev Immunol* 1992, 10:13-49.
45. Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA: **Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories**. *Genome Res* 2009, **19**(5):770-777.
46. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order**. *Nat Rev Genet* 2004, **5**(4):299-310.
47. Sproul D, Gilbert N, Bickmore WA: **The role of chromatin structure in regulating the expression of clustered genes**. *Nat Rev Genet* 2005, **6**(10):775-781.
48. Michalak P: **Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes**. *Genomics* 2008, **91**(3):243-248.
49. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA et al: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains**. *Science* 2001, **291**(5507):1289-1292.
50. Gierman HJ, Indemans MH, Koster J, Goetze S, Seppen J, Geerts D, van Driel R, Versteeg R: **Domain-wide regulation of gene expression in the human genome**. *Genome Res* 2007, **17**(9):1286-1295.

51. Chakalova L, Debrand E, Mitchell JA, Osborne CS, Fraser P: **Replication and transcription: shaping the landscape of the genome**. *Nat Rev Genet* 2005, **6**(9):669-677.
52. Croft JA, Bridger JM, Boyle S, Perry P, Teague P, Bickmore WA: **Differences in the localization and morphology of chromosomes in the human nucleus**. *J Cell Biol* 1999, **145**(6):1119-1131.
53. Habermann FA, Cremer M, Walter J, Kreth G, von Hase J, Bauer K, Wienberg J, Cremer C, Cremer T, Solovei I: **Arrangements of macro- and microchromosomes in chicken cells**. *Chromosome Res* 2001, **9**(7):569-584.
54. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E: **A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly**. *Human molecular genetics* 2003, **12**(14):1725-1735.
55. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM: **Scanning human gene deserts for long-range enhancers**. *Science* 2003, **302**(5644):413.
56. Qin Y, Kong LK, Poirier C, Truong C, Overbeek PA, Bishop CE: **Long-range activation of Sox9 in Odd Sex (Ods) mice**. *Human molecular genetics* 2004, **13**(12):1213-1218.
57. West AG, Fraser P: **Remote control of gene transcription**. *Human molecular genetics* 2005, **14 Spec No 1**:R101-111.
58. Tetko IV, Haberer G, Rudd S, Meyers B, Mewes HW, Mayer KF: **Spatiotemporal expression control correlates with intragenic scaffold matrix attachment regions (S/MARs) in Arabidopsis thaliana**. *PLoS computational biology* 2006, **2**(3):e21.
59. R Development Core Team. R: *A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. 2008..
60. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.

Chapter 2

An introduction to two-color microarray data analysis

Chapter 2 is a summary of the findings and results described in the following six papers:

1. de Koning DJ, Jaffrezic F, Lund MS, Watson M, Channing C, Hulsegge I, Pool MH, Buitenhuis B, Hedegaard J, Hornshoj H, Jiang L, Sorensen P, Marot G, Delmas C, Le Cao KA, San Cristobal M, Baron MD, Malinverni R, Stella A, Brunner RM, Seyfert HM, Jensen K, Mouzaki D, Waddington D, Jimenez-Marin A, Perez-Alegre M, Perez-Reinado E, Closset R, Detilleux JC, Dovc P, Lavric M, Nie H, Janss L: **The EADGENE Microarray Data Analysis Workshop (open access publication)**. *Genet Sel Evol* 2007, **39**(6):621-631.

2. Jaffrezic F, de Koning DJ, Boettcher PJ, Bonnet A, Buitenhuis B, Closset R, Dejean S, Delmas C, Detilleux JC, Dovc P, Duval M, Foulley JL, Hedegaard J, Hornshoj H, Hulsegge I, Janss L, Jensen K, Jiang L, Lavric M, Le Cao KA, Lund MS, Malinverni R, Marot G, Nie H, Petzl W, Pool MH, Robert-Granie C, San Cristobal M, van Schothorst EM, Schuberth HJ, Sorensen P, Stella A, Tosser-Klopp G, Waddington D, Watson M, Yang W, Zerbe H, Seyfert HM: **Analysis of the real EADGENE data set: comparison of methods and guidelines for data normalisation and selection of differentially expressed genes (open access publication)**. *Genet Sel Evol* 2007, **39**(6):633-650.

3. Sorensen P, Bonnet A, Buitenhuis B, Closset R, Dejean S, Delmas C, Duval M, Glass L, Hedegaard J, Hornshoj H, Hulsegge I, Jaffrezic F, Jensen K, Jiang L, de Koning DJ, Le Cao KA, Nie H, Petzl W, Pool MH, Robert-Granie C, San Cristobal M, Lund MS, van Schothorst EM, Schuberth HJ, Seyfert HM, Tosser-Klopp G, Waddington D, Watson M, Yang W, Zerbe H: **Analysis of the real EADGENE data set: multivariate approaches and post analysis (open access publication)**. *Genet Sel Evol* 2007, **39**(6):651-668.

4. Neerincx PB, Rauwerda H, Nie H, Groenen MA, Breit TM, Leunissen JA: **OligoRAP - an Oligo Re-Annotation Pipeline to improve annotation and estimate target specificity**. *BMC proceedings* 2009, **3 Suppl 4**:S4.

5. Neerincx PB, Casel P, Prickett D, Nie H, Watson M, Leunissen JA, Groenen MA, Klopp C: **Comparison of three microarray probe annotation pipelines: differences in strategies and their effect on downstream analysis**. *BMC proceedings* 2009, **3 Suppl 4**:S1.

6. Hedegaard J, Arce C, Bicciato S, Bonnet A, Buitenhuis B, Collado-Romero M, Conley LN, Sancristobal M, Ferrari F, Garrido JJ, Groenen MA, Hornshoj H, Hulsegge I, Jiang L, Jimenez-Marin A, Kommadath A, Lagarrigue S, Leunissen JA, Liaubet L, Neerincx PB, Nie H, Poel J, Prickett D, Ramirez-Boo M, Rebel JM, Robert-Granie C, Skarman A, Smits MA, Sorensen P, Tosser-Klopp G, Watson M: **Methods for interpreting lists of affected genes obtained in a DNA microarray experiment**. *BMC proceedings* 2009, **3 Suppl 4**:S5.

2.1 Introduction

Genomics involves the analysis of large datasets obtained from various biological experiments. One type of large-scale experiment involves monitoring the expression levels of thousands of genes simultaneously under particular conditions, often referred to as expression profiling or gene expression analysis. Microarray technology has become one of the indispensable tools to monitor genome wide expression levels of genes in a given organism. A microarray is typically a glass slide on to which DNA molecules (often called probes) are fixed in an orderly manner at specific locations called spots. A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. Microarray technology makes surveying genome-wide gene expression in an organism possible and the quantity of data generated from each experiment is enormous. This chapter briefly introduces the basic statistical processes needed to process the data derived from a microarray experiment, including background correction, single-array normalization, and multi-array normalization. Several normalization methods will be summarized and discussed in the following parts of this chapter. One of the available functional analysis methods to identify significantly enriched biological pathways/functions in the gene list of interest will also be introduced after the normalization steps.

Since this PhD project is part of EADGENE (European Animal Disease Genomic Network of Excellence) network (www.eadgene.info), and microarray has been used as one of the most popular techniques for transcriptomic studies in EADGENE network, the statistical/bioinformatics' analysis of microarray data has been one of the major concerns for the network. In 2007 and 2008, two workshops focusing on microarray data normalization and post-analysis of microarray data were organized, respectively, aimed at comparing several different software and analysis methods on the same microarray dataset to see the different effects of different methodologies on both microarray normalizations [1] and functional analysis after the normalization [2]. In this chapter, some of the key findings from the two EADGENE workshops which have been published in two series of papers, in Genetics Selection Evolution [3, 4] and in BMC Proceedings [5-7], will be summarized.

Given the fact that the genome information of most farm animal species are far from complete, the available genome information for these species in the current genome databases evolves relatively fast. The existing microarray platforms (probe designs) for, in this case, chickens are lagging behind the current genome information available in the updated databases. In this chapter, we will also describe a bioinformatics tool that can be used to update probe annotations based on the newest genome information available using sequence information of the probes. By doing so, the most updated and accurate probe annotations for the microarray platforms are available and this allows a more reliable biological interpretation of the microarray experiments.

In this thesis, we carried out two genome-wide gene expression survey in several chicken tissues in different stages (adult and embryonic stage), the number of tissues was large in each experiment, therefore, we used the common reference design for both experiment in this thesis. \the common reference design makes the hybridization scheme and data analysis easier when the number of conditions involved is larger. The design of a microarray experiment depends on the biological question to be addressed, this aspect has been discussed in detail in a number of papers [8-11], and therefore design issues will not be discussed in this chapter.

2.2. Common biological questions of microarray experiments

Key questions that in general are addressed within a microarray experiment are:

- Which genes are differentially expressed between two conditions or among several different conditions?
- Which genes are co-regulated under a set of conditions?
- Which genes are co-regulated by a common transcription factor?
- Which samples are more similar to each other according to global gene expression patterns
- To understand genomic architecture by studying transcriptome.

The microarray is used to obtain a rough estimate of the relative amount of RNA molecules between two samples for each gene. Often (e.g. in case of tissues) an average for a large number of different cell types is obtained. A number of problems need to be considered while using microarrays such as the reproducibility of the results and the number of replicates needed. A statistical analysis of the data is consequently performed in order to make use of the data and interpret the microarray data into biology.

2.3 Microarray data analysis

Microarray analyses have become an important tool in animal genomics. While their use is becoming widespread, there are still many questions regarding the design of the experiment as well as the best way to analysis the data. Bioconductor [12] developed in R [13] has become popular for microarray data analysis, because it is an open source program with many different available statistical algorithms dealing with microarray data normalization, differential expression identification, clustering, pathway analysis and some other bioinformatics tool querying online databases. In this thesis, all the microarray data analyses were performed using the bioconductor packages within R.

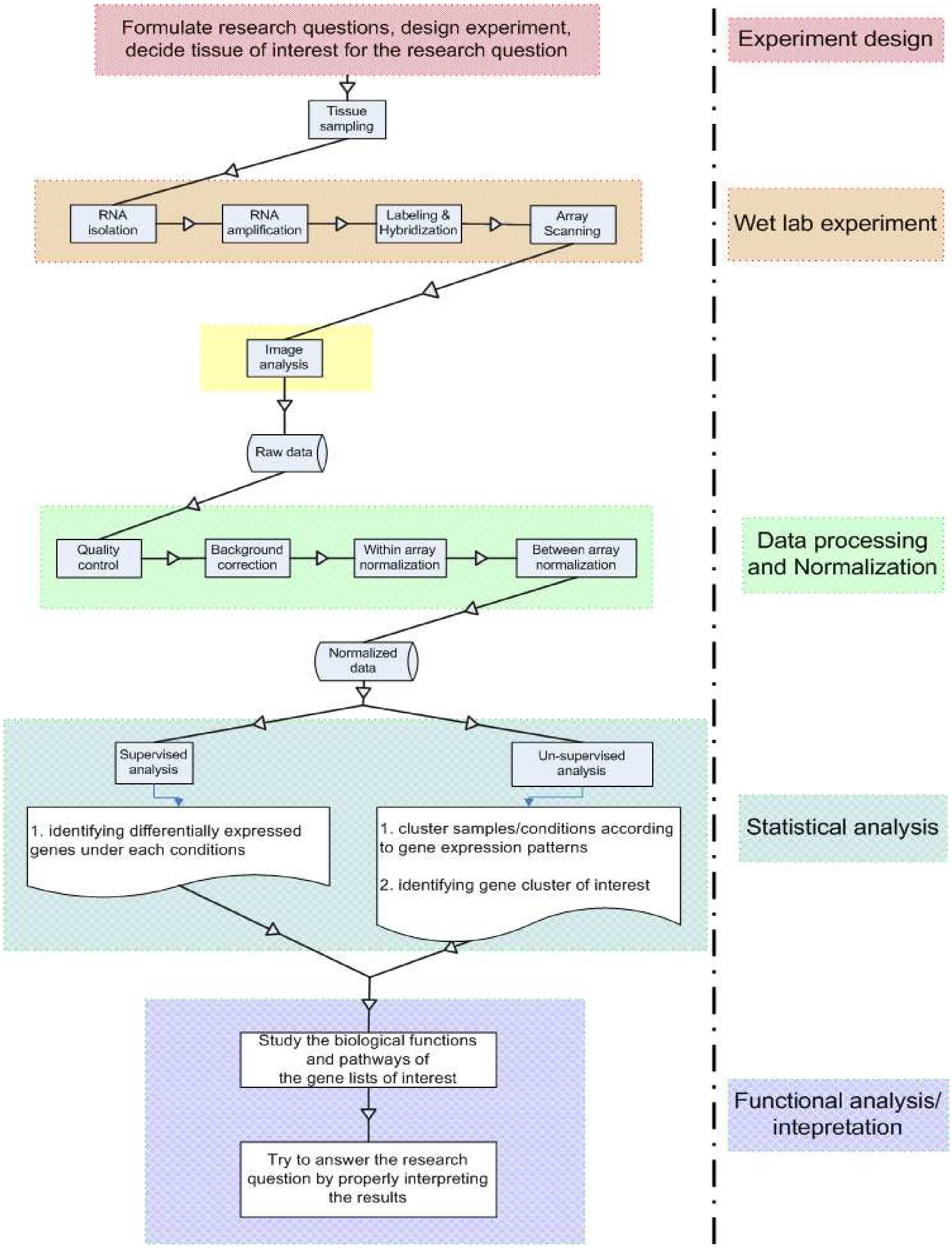

Formulate research questions, design experiment, decide tissue of interest for the research question
Experiment design
Tissue sampling
RNA isolation
RNA amplification
Labeling & Hybridization
Array Scanning
Wet lab experiment
Image analysis
Raw data
Quality control
Background correction
Within array normalization
Between array normalization
Data processing and Normalization
Normalized data
Supervised analysis
Un-supervised analysis
1. identifying differentially expressed genes under each conditions
1. cluster samples/conditions according to gene expression patterns
2. identifying gene cluster of interest
Statistical analysis
Study the biological functions and pathways of the gene lists of interest
Try to answer the research question by properly interpreting the results
Functional analysis/ intepretation

Figure 1. An overview of microarray data analysis: from wet lab experiment to down-stream analysis.

Microarray data analysis generally includes several different steps, data processing and normalization, statistical analysis, and functional analysis (interpretation of the gene list). An overview of microarray data analysis is shown in Figure 1.

2.3.1 Data normalization

After quantification of the scanned image files of each slide, the raw data needs to be further processed to be suitable for any downstream statistical analysis. Here we introduce several standard microarray data processing and normalization methods available in the R package Limma [14] for two-color microarrays: a) background correction; b) within array normalization; c) between array normalization.

a) Background correction:

On a microarray slide, the measured fluorescence intensity of any spot is a combination of the background intensity around the spot and the intensity from the hybridization level of the mRNA samples to the spotted DNA. Background fluorescence can arise from several sources, such as non-specific binding of labeled sample to the array surface and processing effects such as deposits left after the wash stage or optical noise from the scanner. Removal of ambient, non-specific signal from the total intensity is known as ‘background correction’. Background correction is necessary to estimate the true hybridization level of the cDNA.

Most image analysis software packages (e.g., GenePix) provide estimates for the intensity for the "foreground" and "background" of two channels for every spot. The common approach to further analyze such data is to first subtract the background from the foreground for each channel and to use the ratio of these two results as the estimate of the expression level. This approach may cause problems when the foreground intensity is smaller than the background intensity for a channel of a spot, because of the log2 transformation, that spot yields no usable data. Several different background correction methods are available, for example, Ritchie et al. [15] summarized eight common background correction methods (i.e. Standard, Kooperberg, Edwards, Normexp, Normexp+offset, Vsn, Morph, and No background correction) and compared effects of using different methods for background correction. They concluded that the best performance was achieved by normexp + offset whereas the standard method of background subtraction is the worst method [15].

The normexp+offset method has been employed as the background correction method for data analysis in this thesis.

b) Within array normalization:

Generally, microarray data are relatively noisy, even within a single array the log-ratios will likely depend on the intensity, so the distribution will show the artifacts and not the regulation of genes. A MA plot is used to visualize intensity-dependent ratio of raw microarray data (Figure 2 (a)). The MA plot uses M as the y-axis and A as the x-axis. The MA plot gives a quick overview of the data. MA values are defined as follows for each probe:

$$M = log_2R - log_2G$$

$$A = \frac{1}{2} \times (log_2R + log_2G)$$

A typical M-A plot of a two-color microarray would show a “banana” shape (Figure 2 (a)), this indicates that the ratios (M) are dependent on the intensities (A), especially at the lower intensities. After “Background correction” described in previous section, all the negative control spots (highlighted in yellow in Figure 2) shrank towards the lower end and surrounded around M=0 line (Figure 2(b)). After within-array normalization, the majority of the spots on the array were distributed along the M=0 (log2(1)=0) line in a, more or less, symmetrical pattern above and below M=0 line (Figure 2(c)). All the negative controls (yellow) are distributed along M=0 at the low intensity levels, and all the positive controls (highlighted in red) are distributed along M=0 at different intensity levels, both (yellow and red spots) indicate that the normalization processes work well.

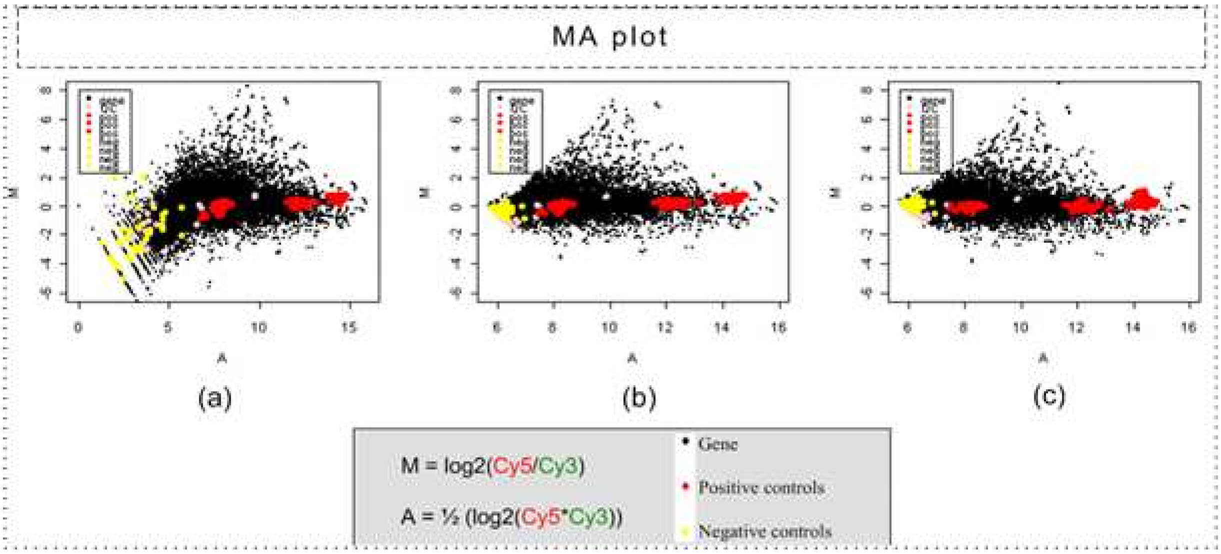


Figure 2. MA-plots of a single array: (a) before normalization, (b) after background correction, (c) after within array normalization.

The major assumptions for the normalization are as listed below and if one or more assumptions are violated, the normalization might lead to wrong results: (1) The majority of the genes are not differentially expressed (M value around 0); (2) The number of genes up- and down-regulated is small and approximately equal. This is not true for arrays with selected genes, but is true for most genome-wide expression arrays. (3) The genes are expressed at a wide range of total intensity (A value). This may not be true for conditions that are extremely different.

Normalization of the data within an array is a two-step process including a correction for spatial bias, and a correction for intensity-dependent bias. Correction for spatial bias is usually carried out separately for each block (print-tip) of each array by either subtracting the median for each block or by subtracting the corresponding row and column means [16]. The intensity dependent bias is removed by either print-tip loess correction [17], or by a global Loess correction [18]. The print-tip Loess is a commonly used within array normalization method, which is available within Limma. It removes the spatial and intensity-dependent bias within each array. No major differences were found when comparing global Loess and print-tip Loess [3]. Therefore, in this thesis, print-tip loess has been employed as the method for within-array normalization to correct the spatial and intensity-dependent bias within each array.

c) Between array normalization:

Probes/genes on different replicated arrays are not comparable before multi-array normalization or between array normalization, because the existence of random and systematical errors between different individual arrays. Data normalized within each individual array need to be further normalized between multi-arrays. In Limma, several options are available for between array normalization, i.e. "scale", "quantile", "Aquantile", "Gquantile", "Rquantile", "Tquantile" or "vsn". The choice of between array normalization methods depends on the biological assumptions made for the experiment. For example, “quantile” is a good option when comparing two groups of samples from the same tissue treated differently and no large proportion of genes is expected to be differentially expressed between the two conditions. When using a common reference design scheme, “Rquantile” or “Gquantile” are good options depending on how the reference sample is labeled. In this thesis, we employed the common reference design for our array experimental design, and the reference samples were always labeled with Cy5 (Red), so “Rquantile” was employed as between array normalization method for this thesis. Since the reference sample is identical for different individual arrays, forcing the identical data distribution of the reference sample (red) across all arrays will enable us to compare different arrays within each experiment (Figure 3).

Jaffrezic et al. [3] concluded that the normalization process is important for dealing with replicated experiments. It relies on prior assumptions which, if they are violated, lead to incorrect results. The Loess function is often a safe choice for the normalization even if it can become unstable at the left end of the data on a MA plot, i.e. those genes with very low expression in both channels. At last the visual inspection is necessary by visualizing data using different plots before and after each normalization step.

Last but not least, normalization allows us to compare data from one array to another, normalization is used to remove signals that might obscure biological information. However, the process of normalization is likely to remove some of the biological information as well. Therefore, choosing the optimal normalization methods based on specific microarray data is essential for a successful interpretation of microarray data.

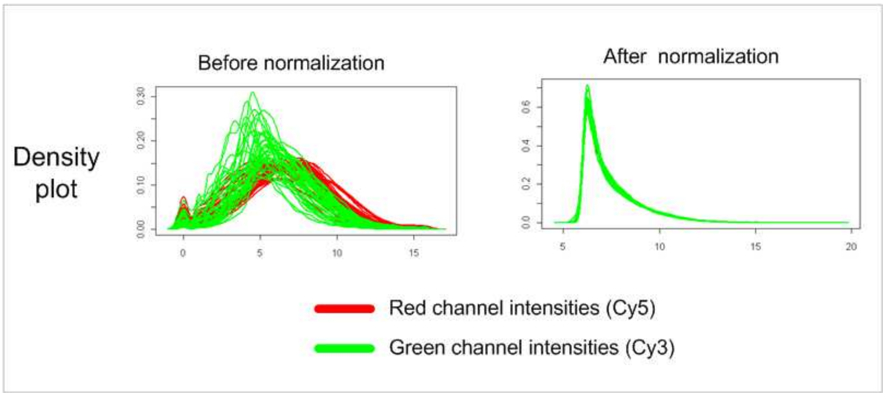


Figure 3. Density plots of multiple arrays before normalization, and after normalization (print-tip loess + Rquantile).

2.3.2 Identification of differentially expressed genes

Traditionally, differentially expressed genes are inferred by a fixed threshold cut off method (for example a two fold increase or decrease). However this is statistically inefficient, the main reason being that there are numerous systematic and biological variations that occur during a microarray experiment. Although some of the systematic variations such as dye bias can be effectively removed by normalization, random biological variations and physiological variations are more difficult to handle. Because of these underlying variations, merely using a fixed threshold to infer significance might increase the proportion of false positives or false negatives. A better framework of significance inference includes calculation of a statistic based on replicate array data from ranking genes according to their possibilities of differential expression and selection of a cut-off value from rejecting the null-hypothesis that the gene is not differentially expressed. Setting a cut-off for differential expression is difficult because one has to balance the false positives and false negatives. Furthermore, performing statistical tests for tens of thousands of genes creates a multiple hypothesis-testing problem. So a p-value of 0.05 is likely to exaggerate false positives. As it is often acceptable to have a few false positives if the majority of true positives are chosen, it might be therefore more practical to control the false discovery rate (FDR) which is the expected proportion of false positives among the number of rejected hypotheses.

During The EADGENE Microarray Data Analysis Workshop [1], several different methods were also applied to the array data after normalization [3, 4]. Jaffrezic et al. [3] discussed issues about identifying differentially expressed genes and multiple testing problems. They showed that, for the identification of the differentially expressed genes, the method implemented in the Bioconductor package Limma was prefered. This method allows complex designs and provides robust t- and F-statistics for differential gene expression by usingempirical Bayes methods (eBayes) for shrinking the residual variances of genes towards their approximate median value. This approach is based on an inverse chi-square prior on the variances [19]. Regarding the correction for multiple tests, the classical Benjamini and Hochberg [20] correction at a 5% False Discovery Rate (FDR) was used as common threshold. Furthermore, Sorensen et al. [4] discussed some post-normalization methods, such as hierarchical clustering (HC), principal component analysis (PCA) for class discovery in the samples and identifying co-expressed genes across different conditions.

2.3.3 From gene lists to biological interpretation

Once genes have been identified that were differentially regulated under certain conditions, or when a cluster of genes has been identified showing interesting expression patterns across a set of conditions, the next phase is to identify the biological processes responsible for these changes. Currently, Gene Ontology [21] and KEGG [22] pathways are two popular choices for gene functional annotation to help to uncover the biological processes involved. The Post-Analysis Workshop [2] organized by EADGENE and SABRE (Cutting Edge Genomics for Sustainable Animal Breeding) in November 2008, focused on the post analysis of microarray data and the usage of these two resources [7]. The participating groups were provided with identical lists of microarray probes, including test statistics for three different contrasts, and the normalized log-ratios for each array, to be used as the starting point for interpreting the affected probes. The tools used by the different groups were: Ingenuity Pathway Analysis, MAPPFinder, Limma, GOstats, GOEAST, GOTM, Globaltest, TopGO, ArrayUnlock, Pathway Studio, GIST and AnnotationDbi. The main focus of the different approaches was to utilize the relation between probes/genes and their gene ontology and pathways to interpret the affected probes/genes. The main results from these analyses showed that the biological interpretation is highly dependent on the statistical method used but that some common biological conclusions can be reached even with very different analysis tools. In chapter 3, a more detailed analysis of interpreting gene list of interest to biology using R packages, biomaRt [23], AnnotationDbi [24] and GOstats [25], is described.

2.4 Re-annotation of the chicken 20K microarray probe set

The microarray platform used in the experiments described in chapters 3-6 of this thesis is the ARK-Genomics Chicken 20 K array [26] consisting of 20.460 unique probes ranging in length from 60 to 75 nucleotides with the majority of the probes being 70 nucleotides long. The array was designed based on chicken genome assembly WASHUC1 (December 2004) including the following information: 1) INSDC (DDBJ/EMBL/GenBank) ESTs/cDNAs including the UMIST ChESTs, 2) Ensembl 30 with gene models based on various sources ranging from highly reliable chicken UniProtKB/Swiss-Prot proteins to relatively unreliable ab initio in silico gene predictions, 3) miRBase micro RNAs. Although, the release of the chicken genome sequence in 2004 [27] has been a landmark for chicken biology, it still is a draft genome sequence leaving much room for further improvement on assembly quality, sequence coverage, and gene discovery. The biological functions of many chicken genes are not known, and the lack of a well-annotated chicken genome did limit the possibilities to fully explore the tools which were being used to uncover biological processes within lists of interesting genes obtained from microarray experiments.

High throughput gene expression studies using oligonucleotide microarrays depend on the specificity of each oligonucleotide (oligo or probe) for its target gene. However, target specific probes can only be designed when a reference genome of the species at hand is completely sequenced, when this genome is completely annotated and when the genetic variation of the sampled individuals is completely known. Unfortunately there is not a single species for which such a complete data set is available. Therefore, it is important that probe annotation is updated frequently for an optimal interpretation of microarray experiments. Neerincx et al. [5] presented their work on oligo reannotation using OligoRAP, a pipleline to automatically update the annotation of oligo libraries and estimate oligo target specificity. OligoRAP uses a reference genome assembly with Ensembl and Entrez Gene annotation supplemented with a set of unmapped transcripts derived from RefSeq and UniGene to handle assembly gaps. OligoRAP produces alignments of each oligo with the reference assembly as well as with unmapped transcripts. These alignments are remapped to the annotation sources, which results in a concise, as complete as possible and up-to-date annotation of the oligo library. Neerincx et al [5] found dramatic differences in the updated annotation and target specificity for the ARK-Genomics 20 K chicken array as compared to the original data, emphasizing the need for regular updates of the probes as well as the annotation of this array platform. In addition to the reannotation platform descibed above, Neerincx et al. [28] made a comparison among three different oligo re-annotation pipelines (IMAD [28], OligoRAP, and sigReannot

[29]) and showed that the differences in updated annotation are mainly due to different thresholds for hybridisation potential filtering of oligo versus target-gene alignments and different policies for expanding annotation using indirect links. Furthermore, the effect of differences in the updated annotation on the functional analysis (GO/KEGG enrichment analysis) was analyzed and the differences in the updated annotation packages had a large effect on GO term enrichment analyses. It was proposed that annotation tools should provide metadata describing the relationships between oligos and the annotation assigned to them. These relationships can then be used to judge the varying degrees of reliability allowing users to fine-tune the balance between reliability and coverage. This is important as it can have a large effect on functional microarray analyses as exemplified by the lack of consensus on almost one third of the terms found with GO term enrichment analysis based on updated IMAD, OligoRAP or sigReannot annotation. It was further concluded that a consensus threshold for probe updating is needed for different re-annotating pipelines to reach more consensus results in functional analyses.

In summary, the array normalization procedure at different steps described above (highlighted with underscore) was used in the data analysis described in the following chapters of this thesis. The updated probe function annotation used in the following analysis was derived from oligoRAP re-annotation pipeline as described by Neerincx et al. [5]. The functional analysis of microarray data introduced in this chapter is further described in detail in Chapter 3.

References

1. de Koning DJ, Jaffrezic F, Lund MS, Watson M, Channing C, Hulsegge I, Pool MH, Buitenhuis B, Hedegaard J, Hornshoj H et al: **The EADGENE Microarray Data Analysis Workshop (open access publication)**. *Genet Sel Evol* 2007, **39**(6):621-631.
2. Jaffrezic F, Hedegaard J, Sancristobal M, Klopp C, de Koning DJ: **The EADGENE and SABRE post-analyses workshop**. *BMC Proc* 2009, **3 Suppl 4**:I1.
3. Jaffrezic F, de Koning DJ, Boettcher PJ, Bonnet A, Buitenhuis B, Closset R, Dejean S, Delmas C, Detilleux JC, Dovc P et al: **Analysis of the real EADGENE data set: comparison of methods and guidelines for data normalisation and selection of differentially expressed genes (open access publication)**. *Genet Sel Evol* 2007, **39**(6):633-650.
4. Sorensen P, Bonnet A, Buitenhuis B, Closset R, Dejean S, Delmas C, Duval M, Glass L, Hedegaard J, Hornshoj H et al: **Analysis of the real EADGENE data set: multivariate approaches and post analysis (open access publication)**. *Genet Sel Evol* 2007,

39(6):651-668.

5. Neerincx PB, Rauwerda H, Nie H, Groenen MA, Breit TM, Leunissen JA: **OligoRAP - an Oligo Re-Annotation Pipeline to improve annotation and estimate target specificity**. *BMC proceedings* 2009, *3 Suppl 4*:S4.
6. Neerincx PB, Casel P, Prickett D, Nie H, Watson M, Leunissen JA, Groenen MA, Klopp C: **Comparison of three microarray probe annotation pipelines: differences in strategies and their effect on downstream analysis**. *BMC proceedings* 2009, **3 Suppl 4**:S1.
7. Hedegaard J, Arce C, Bicciato S, Bonnet A, Buitenhuis B, Collado-Romero M, Conley LN, Sancristobal M, Ferrari F, Garrido JJ et al: **Methods for interpreting lists of affected genes obtained in a DNA microarray experiment**. *BMC Proc* 2009, **3 Suppl 4**:S5.
8. Kerr MK, Churchill GA: **Experimental design for gene expression microarrays**. *Biostatistics* 2001, **2**(2):183-201.
9. Yoo C, Cooper GF: **A computer-based microarray experiment design-system for gene-regulation pathway discovery**. *AMIA Annu Symp Proc* 2003:733-737.
10. Lee KM, Kim JH, Kang D: **Design issues in toxicogenomics using DNA microarray experiment**. *Toxicol Appl Pharmacol* 2005, **207**(2 Suppl):200-208.
11. Yoo C, Cooper GF, Schmidt M: **A control study to evaluate a computer-based microarray experiment design recommendation system for gene-regulation pathways discovery**. *J Biomed Inform* 2006, **39**(2):126-146.
12. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.
13. R Development Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. 2008..
14. Smyth GK: **Limma: Linear models for microarray data**. In *Stat Appl Genet Mol Biol. Volume 3.* New York: Springer; 2004.
15. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK: **A comparison of background correction methods for two-colour microarrays**. *Bioinformatics* 2007, **23**(20):2700-2707.
16. Baird D, Johnstone P, Wilson T: **Normalization of microarray data using a spatial mixed model analysis which includes splines**. *Bioinformatics* 2004, **20**(17):3196-3205.
17. Smyth GK, Speed T: **Normalization of cDNA microarray data**. *Methods* 2003, **31**(4):265-273.

18. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation**. *Nucleic Acids Res* 2002, **30**(4):e15.

19. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article 3.

20. Benjamini Y and Hochberg Y: Controlling the false discovery rate: **A practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society* 1995, **Series B 57**:289–300.

21. **Gene Ontology** [http://www.geneontology.org/]

22. **Kegg pathway database** [http://www.genome.jp/kegg/pathway.html]

23. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis**. *Bioinformatics* 2005, **21**(16):3439-3440.

24. Pages H, Carlson M, Falcon S and Li N: **AnnotationDbi: Annotation Database Interface**. 2008, package version 1.4.0.

25. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association**. *Bioinformatics* 2007, **23**(2):257-258.

26. **ARK-Genomics Chicken 20K Oligo Array** [http://www.arkgenomics.org/microarrays/]

27. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.

28. Prickett D, Watson M: **IMAD: flexible annotation of microarray sequences**. *BMC Proc* 2009, **3 Suppl 4**:S2.

29. Casel P, Moreews F, Lagarrigue S, Klopp C: **sigReannot: an oligo-set re-annotation pipeline based on similarities with the Ensembl transcripts and Unigene clusters**. *BMC Proc* 2009, **3 Suppl 4**:S3.

Chapter 3

Microarray data mining using Bioconductor packages

Haisheng Nie[1], Pieter B.T. Neerincx[2], Jan van der Poel[1], Francesco Ferrari[3], Silvio Bicciato[4], Jack A.M. Leunissen[2], Martien A.M. Groenen[1]

1. Animal Breeding and Genomics Centre
   Wageningen University
   The Netherlands
2. Laboratory of Bioinformatics
   Wageningen University
   The Netherlands
3. Department of Biology
   University of Padova
   Italy
4. Department of Biomedical Sciences
   University of Modena and Reggio Emilia
   Italy

Based on: Nie et al. *BMC Proceedings* 2009, 3(Suppl 4): S9

Abstract

Background

Eimeria are obligate intracellular protozoan parasites which can affect chickens and continuous exposure to Eimeria can result in protective immunity. The process leading to protective immunity was investigated by studying the host reactions after homologous or heterologous secondary infections using microarrays. The array data was used in the EADGENE and SABRE post-analyses workshop, and this paper describes the results of a Gene Ontology (GO) term enrichment analysis of chicken microarray data using the Bioconductor packages. By checking the enriched GO terms of differentially expressed (DE) genes from the microarray data, this analysis aimed to investigate the host reactions in chickens occurring shortly after a secondary challenge with either a homologous or heterologous species of Eimeria. The results of GO enrichment analysis using GO terms annotated to chicken genes and GO terms annotated to chicken-human orthologous genes were also compared. Furthermore, a locally adaptive statistical procedure (LAP) was performed to test differentially expressed chromosomal regions, rather than individual genes, in the chicken genome after Eimeria challenge.

Results

GO enrichment analysis identified significant (raw p-value < 0.05) GO terms for all three contrasts included in the analysis. Some of the GO terms linked to, generally, primary or secondary immune responses indicating the GO enrichment analysis is a useful approach to analyze microarray data. The comparisons of GO enrichment results using chicken gene information and chicken-human orthologous gene information showed more refined GO terms related to immune responses when using chicken-human orthologous gene information, this suggests that using chicken-human orthologous gene information has higher power to detect significant GO terms with more refined functionality. Furthermore, three chromosomal regions were identified to be significantly up-regulated in the contrast MM8-PM8 (q- value < 0.01).

Conclusion

Overall, this paper describes a practical approach to analyze microarray data in farm animals where the genome information is still incomplete. For farm animals, such as chicken, with currently limited gene annotation, borrowing gene annotation information from orthologous genes in well-annotated species, such as human, will help improve the pathway analysis results substantially. Furthermore, LAP analysis approach is a relatively new and very useful way to be applied in microarray analysis.

Background

Eimeria are obligate intracellular protozoan parasites which can affect chickens and continuous exposure to Eimeria can result in protective immunity. The process leading to protective immunity was investigated by studying the host reactions after homologous or heterologous secondary infections. A total of 125 one-day-old Ross 308 male broilers were randomly divided in five groups of 25 broilers each. At 7 days of age, three groups were inoculated with phosphate buffered saline (P) and two groups were inoculated with E. maxima (M). A secondary challenge followed at day 21 of age. This challenge was with PBS (P), E. maxima (M) or with E. acervulina (A), forming five challenge groups PP, PM, PA, MM and MA. Five chickens from each group were killed at 8 and 24 hours after the second challenge and specific regulations of gene expression profiles in the jejunum were monitored using chicken whole genome oligonucleotide microarrays (ARK-Genomics *Gallus gallus* 20 K v1.0). The obtained microarray data was normalised and analysed and lists of affected genes were obtained for different contrasts. The result of the contrasts MM8-PM8, MM8-MA8 and MM8-MM24 were provided for this workshop as three lists including all microarray probes and test statistics for the three different contrasts. The number of affected probes for each contrast is shown in Table 1.

Table 1. The contrasts used in the workshop. The number of significantly (FDR <= 0.05) DE genes for the three different contrasts used in the workshop.

Contrast:	MM8.PM8	MM8.MA8	MM8.MM24
Repressed	803	58	639
Induced	923	23	152

The normalised log-ratios for each array were furthermore used in the workshop. The contrasts address different biological questions: differences between secondary and primary challenge (MM8-PM8), differences between homologous and heterologous challenge (MM8-MA8) and differences between two time points of a homologous challenge (MM8-MM24). The microarray data is available at the ArrayExpress database [1] under accession number E-MEXP-1972 and the three gene lists can be downloaded from supplementary material of Hedegaard et al. [2]

This paper is part of the The EADGENE and SABRE post-analyses workshop [3]. In this analysis, we focus our analysis on the gene lists from three contrasts: MM8-PM8, MM8-MA8 and MM8-MM24. Each contrast has both up- and down-regulated significant gene lists, in total six gene lists were used for Gene Ontology [4] term enrichment analysis.

The analysis in this paper was carried out using a number of different Bioconductor [5] packages (release version: BioC 2.3); GOstats [6], AnnotationDbi [7], and biomaRt [8]. Package Gostats uses hypergeometic test to identify significantly enriched GO terms in gene lists of interest.

Package GOstats also provides conditional hypergeometric test which uses the relationship among GO terms to decorrelate the results. Package AnnotationDbi Provides an interface and database connection code for annotation data packages using SQLite data storage, the annotation data packages were needed for GOstats package. Package biomaRt provides an R interface to BioMart databases [9].

To investigate the effects of different sources of microarray probe annotation on GO term enrichment analysis, two analyses were carried out: one used chicken gene information and the other one used chicken-human orthologous gene information.

Furthermore, a locally adaptive statistical procedure (LAP) [10] was performed to test differentially expressed chromosomal regions, rather than individual genes, in the chicken genome after Eimeria challenge. LAP is a non-parametric model-free statistical method for the identification of differentially expressed chromosomal regions, which accounts for variations in gene distance and density. The method is based on the computation of a standard statistic (e.g. SAM t-statistic) as a measure of the difference in gene expression patterns between groups of samples. The LAP analysis approach is a relatively new and interesting way of analyzing microarray data.

Methods

Chicken 20k oligo array annotation

An updated chicken 20k oligo-array annotation based on Ensembl [11] release 50 was downloaded from EADGENE Oligo Set Annotation Files homepage [12]. Human orthologous genes, if identified, were mapped to the corresponding chicken oligo probes present on the chicken array. The human Ensembl gene IDs were then used to extract human Entrez gene IDs via the Bioconductor package biomaRt by querying to the Ensembl genome database. The resulting human Entrez gene IDs were subsequently used to build a customized chicken array annotation R package using AnnotationDbi.

GO enrichment analysis

A GO term enrichment analysis was carried out using package GOstats and a conditional hypergeometric test algorithm provided within GOstats package was applied to each gene list. The conditional hypergeometric test will identify a GO term as significant if there is evidence beyond that provided by its significant children. The threshold for significance of the hypergeometric test was raw p-values < 0.05. Only GO terms in the category Biological Process (GO_BP) were used in this analysis. Those GO terms were excluded from the result list when Count equal to 1 Or Size equal to 1, i.e. only 1 gene in the DE gene list links to this specific GO term or only 1 gene on the whole array links to this specific GO term.

Differentially expressed chromosomal regions

Differentially expressed chromosomal regions were identified using locally adaptive procedure (LAP). LAP analysis was performed in R [13] and the threshold used in this analysis is q-values < 0.01, where q-value is the false discovery rate calculated from p-values between two group comparisons, i.e. p-values derived from each contrast.

Results and discussion

GO term enrichment analysis

All the GO enrichment analysis results are available in the Additional file 1 and Additional file 2. Here we will focus only on the selected GO terms related to immune response (see Additional file 1) to explain the three contrasts, MM8-PM8, MM8-MA8, and MM8-MM24.

(1) MM8-PM8 contrast

Genes that are up-regulated in the MM8-PM8 contrast show an enrichment of GO terms like, “immune response-activating cell surface receptor signalling pathway”, “proteolysis involved in cellular protein catabolic process” and “focal adhesion formation”. These terms all indicate that the chickens show primary immune responses at 8 hours after PM challenge.

Genes that are down-regulated in the MM8-PM8 contrast show an enrichment of GO terms like, “regulation of B cell differentiation”, “regulation of T cell activation”, “T cell selection” and “regulation of interferon-gamma biosynthetic process”, terms indicative for a secondary immune response at 8 hours after homologous MM challenge.

These results clearly show the induction of different immune responses (primary vs. secondary) in chicken that encountered an Eimeria infection for the first time and chicken that had gone through an Eimeria infection at an earlier time in their life.

(2) MM8-MA8 contrast

No major differences on immune response related GO terms were identified in the MM8-MA8 contrast. These results show that heterologous challenge with MA triggers a very similar immune response as MM. Interestingly, the genes up-regulated in the MM8-MA8 contrast show an enrichment of GO term like “cell death” and “apoptosis”, suggesting that the heterologous challenge caused more severe lesions in the chickens as compared to a homologous challenge.

No evidence is seen that MM8 and MA8 trigger different immune responses in chicken, although the enriched GO terms indicate a more severe pathogenesis in case of heterologous challenge.

(3) MM8-MM24 contrast

As described in the MM8-PM8 contrast result, the homologous challenge already triggered a secondary immune response at 8 hours. No significant GO terms related to secondary immune response were found in MM8-MM24 contrast. The up-regulated genes in MM8-MM24 have enriched GO terms like “positive regulation of NF-kappaB transcription factor activity”, and the down-regulated genes in MM8-MM24 have enriched GO terms like, “T cell receptor signalling pathway” and “interleukin-2 production”. NF-kappaB is a key regulator of several important immune-related pathways and this suggests that immune response activators were already highly up-regulated at 8 hours compared to 24 hours and that a secondary immune responses kept on increasing from 8 hours to 24 hours after homologous challenge with MM.

Multiple testing problems

We have applied “BH” FDR control method for correction for multiple testing using R package multtest [14] and found only a few significant GO terms after correction (data not shown). In this analysis we used threshold of raw p-value < 0.05, the major reasons of not using the FDR control methods are (a) the structure of the GO graph is in conflict with the assumption of independence for the test and (b) multiple testing correction methods do not change the overall ranks of the results, using raw p-value at cut-off would still identify the relative important GO terms in the results.

Annotation Sources comparison

In this section, GO enrichment analysis results using chicken gene annotation and chicken-human orthologous gene annotation are compared. All the GO term enrichment analysis results of this comparison are available in the Additional file 2 and Additional file 3.
The overlap of the results of the GO term enrichment analysis using the chicken gene

information and using the chicken-human orthologous gene information is shown in Figure 1. The overlap of the significant GO terms identified by both annotation sources is limited. Enriched GO terms using chicken genes and using chicken-human orthologous genes, as described above, gave a reasonably good insight of the underlying biological processes in this experiment. The enriched GO terms based on the chicken annotation directly didn't reveal much detail in the ongoing processes after either homologous challenge or heterologous challenge (see Additional file 2). The enriched GO term using the chicken-human orthologous gene information had a higher power to detect significant GO terms (see Additional file 3), which can be explained by the higher coverage of annotation (GO terms) using this approach.

Performing the GO enrichment analysis using chicken-human orthologous genes, on one hand, extensively increased the coverage of the gene annotation of this chicken oligo array platform. Consequently, this increases the power of the hypergeometric test by having more annotated genes in the DE gene lists. On the other hand, care has to be taken by using this approach, as human and chicken are evolutionarily far apart. Therefore, some of the chicken-specific immune response processes may not be identified using this approach. Nevertheless, this approach helps researchers working with farm animals, e.g. chicken, to increase the biological insight from their microarray data by using human orthologous gene information.

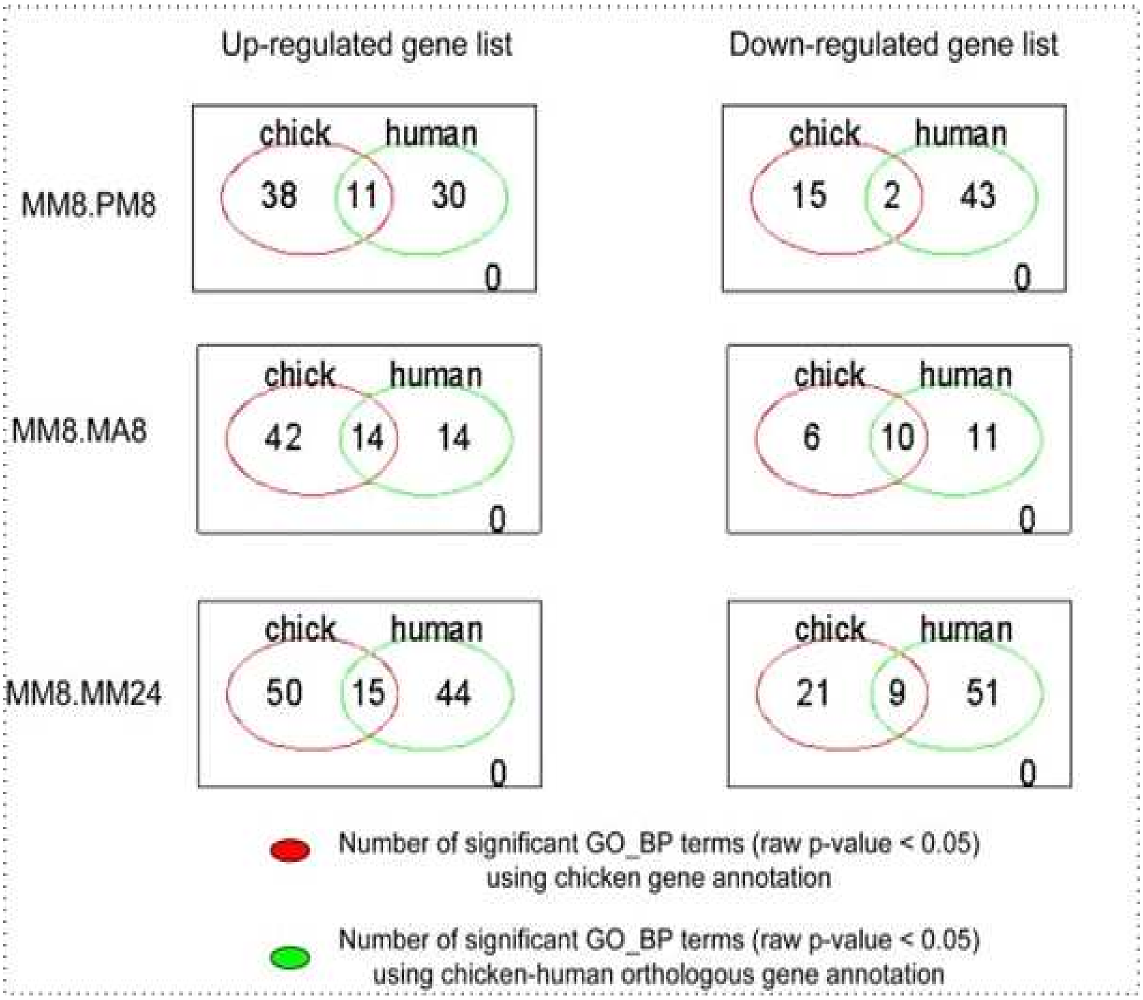

Up-regulated gene list
Down-regulated gene list
MM8.PM8
chick
human
38
11
30
0
chick
human
15
2
43
0
MM8.MA8
chick
human
42
14
14
0
chick
human
6
10
11
0
MM8.MM24
chick
human
50
15
44
0
chick
human
21
9
51
0
Number of significant GO_BP terms (raw p-value < 0.05) using chicken gene annotation
Number of significant GO_BP terms (raw p-value < 0.05) using chicken-human orthologous gene annotation

Figure 1. Comparison of GO term enrichment analysis results: overlap of significantly enriched GO terms (raw p-value < 0.05) between the uses of chicken gene information versus chicken-human orthologous gene information.

Differentially expressed chromosomal regions

Instead of testing enrichment of GO terms, chromosomal locations could be used as “annotation” to test whether certain chromosomal locations are more actively expressed than other regions. In this analysis, the differentially expressed chromosomal locations were identified using locally adaptive procedure (LAP). In total, three significant regions were up-regulated and one region was down-regulated comparing PM and MM infections (see details of those regions in Figure 2 and Additional file 4). No significant regions were identified in other contrasts. The identified differentially expressed chromosomal regions indicate that some of the co-localized genes are co-regulated during homologous challenge by MM, this region-wide gene expression regulation mechanism was reported in several other species [15, 16].

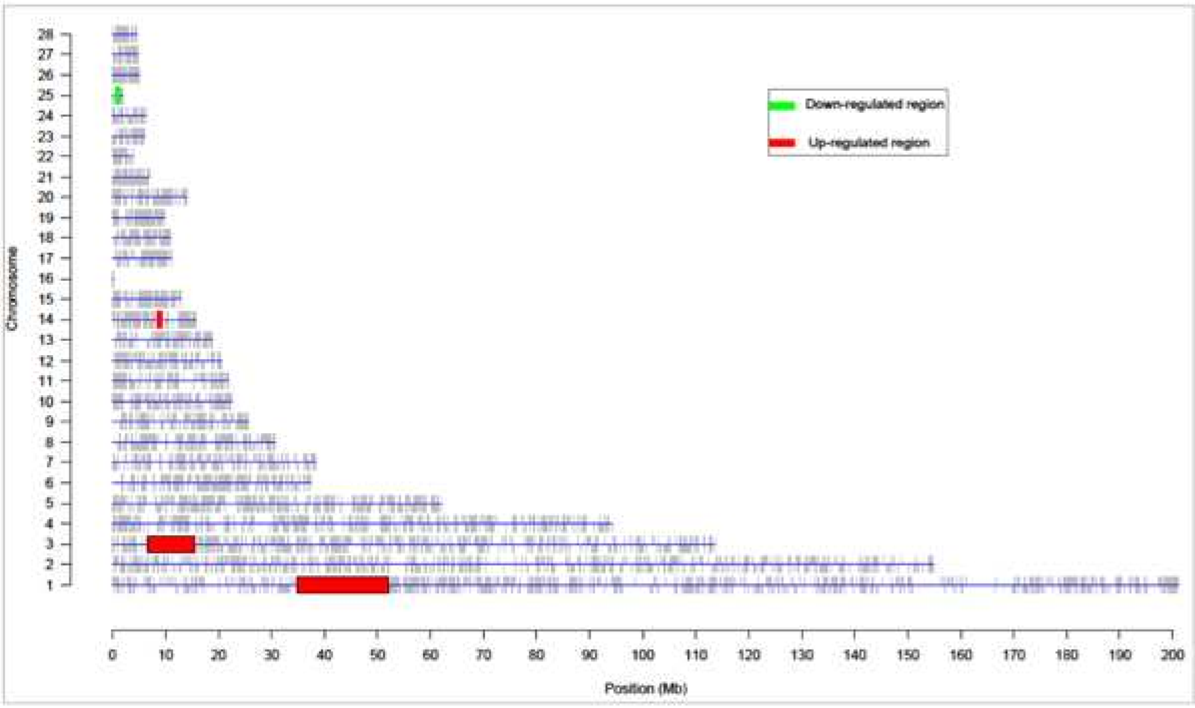


Figure 2. Differentially expressed chromosomal regions for contrast MM8-PM8. This figure showed the differentially expressed chromosomal regions for MM8.PM8 contrast (q-value < 0.01). In total three regions were up-regulated and one region was down-regulated. Red showed the up-regulated chromosomal regions, and Green showed the down-regulated regions.

Conclusion

The GO term enrichment analysis provided a good insight in the biological processes involved in the Eimeria infection experiments. The GO enrichment analysis using several bioconductor packages described in this paper provides a practical, yet powerful, way of analyzing microarray data. Furthermore, the results suggest that using chicken-human orthologous gene information provides better insight in the biological processes underlying this specific microarray experiment than by using the annotation of chicken genes alone. This approach will be a helpful general method for researchers working with microarray data in species with less well annotated-genomes, like those of farm animals.

Furthermore, LAP analysis approach is a relatively new and very useful way to be applied in microarray analysis to identify differentially expressed chromosomal regions under specific experimental conditions.

Abbreviations

DE: Differentially Expressed
FDR: False Discovery Rate
GO: Gene Ontology
GO_BP: Gene Ontology Biological Process
PM: PBS- E. Maxima
MM: E. maxima- E. Maxima
MA: E. maxima -E. acervulina
LAP: locally adaptive statistical procedure

Competing interests

The authors declare that they have no competing interests.

Authors'contributions

HN analyzed the data and drafted the manuscript, all other authors helped to improve the manuscript. PBTN and JAML helped with the re-annotation of the array, JP and MAMG contributed to the biological interpretation of the results, FF and SB performed the analysis of differentially expressed chromosomal regions. All authors read and approved the final manuscript.

Acknowledgements

This work was funded by the EADGENE network. The authors wish to acknowledge Caroline Channing and the other organizers for organizing the workshop, Christophe Klopp and Pierrot Casel from Institut National de la Recherche Agrinomique (INRA) for re-annotating the chicken 20k oligo array, Dr. Annemarie Rebel and colleagues from Animal Sciences Group in Lelystad for providing the microarray data from the chicken infection experiment.

References

1. ArrayExpress [http://www.ebi.ac.uk/microarray-as/ae/]

2. Hedegaard J, Arce C, Bicciato S, Bonnet A, Buitenhuis B, Collado-Romero M, Conley LN, Sancristobal M, Ferrari F, Garrido JJ et al: **Methods for interpreting lists of affected genes obtained in a DNA microarray experiment**. *BMC Proc* 2009, **3 Suppl 4**:S5. [http://www.biomedcentral.com/content/supplementary/1753-6561-3-s4-s5-s1.xls]
3. Jaffrezic F, Hedegaard J, Sancristobal M, Klopp C, de Koning DJ: **The EADGENE and SABRE post-analyses workshop**. *BMC proceedings* 2009, **3 Suppl 4**:I1.
4. **Gene Ontology** [http://www.geneontology.org/]
5. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.
6. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association**. *Bioinformatics* 2007, **23**(2):257-258.
7. Pages H, Carlson M, Falcon S and Li N: **AnnotationDbi: Annotation Database Interface**. R package version 1.4.0.
8. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis**. *Bioinformatics* 2005, **21**(16):3439-3440.
9. **BioMart databases** [http://www.biomart.org]
10. Callegaro A, Basso D, Bicciato S: **A locally adaptive statistical procedure (LAP) to identify differentially expressed chromosomal regions**. *Bioinformatics* 2006, **22**(21):2658-2666.
11. **Ensembl Genome Database** [http://www.ensembl.org]
12. **EADGENE Oligo Set Annotation Files** [http://www.eadgene.info/TheProject/Integration/BiologicalresourcesandfacilitiesWP11/EADGENEOligoSetsAnnotationFiles/tabid/324/Default.aspx]
13. R Development Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. 2008.
14. Pollard KS, Ge Y, Taylor S and Dudoit S: **multtest: Resampling-based multiple hypothesis testing**. R package version 1.22.0.
15. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the Drosophila genome**. *J Biol* 2002, **1**(1):5.
16. Gierman HJ, Indemans MH, Koster J, Goetze S, Seppen J, Geerts D, van Driel R, Versteeg R: **Domain-wide regulation of gene expression in the human genome**. *Genome Res* 2007, **17**(9):1286-1295.

Additional files

Additional file 1 - GO enrichment analysis results with selected immune related GO terms

This table shows GO enrichment results with selected GO_BP terms. (For contrasts MM8.PM8 and MM8.MM24 results, only immune-related GO_BP terms which have at least two genes linked to each one of them were included).

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712752/bin/1753-6561-3-S4-S9-S1.xls

Additional file 2 - GO term enrichment results (raw p-value <0.05) using chicken genes

This table shows the GO enrichment analysis results using chicken gene information.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712752/bin/1753-6561-3-S4-S9-S2.xls

Additional file 3 - GO term enrichment results (raw p-value < 0.05) using chicken-human orthologous genes

This table shows the GO term enrichment analysis results using chicken-human orthologous genes information.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712752/bin/1753-6561-3-S4-S9-S3.xls

Additional file 4 - Differentially expressed chromosomal regions for MM8-PM8 contrast

This table shows the chromosomal locations of three up-regulated chromosomal regions and one down-regulated chromosomal region for MM8-PM8 contrast.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712752/bin/1753-6561-3-S4-S9-S4.xls

Chapter 4

A genome-wide gene expression survey in eight chicken tissues

Haisheng Nie[1], Richard P.M.A. Crooijmans[1], Aart Lammers[2], Evert M. van Schothorst[3], Jaap Keijer[3], Pieter B.T. Neerincx[4], Jack A.M. Leunissen[4], Hendrik-Jan Megens[1] and Martien A.M. Groenen[1]

1. Animal Breeding and Genomics Centre
   Wageningen University
   The Netherlands.

2. Adaptation Physiology Group
   Wageningen University
   The Netherlands

3. Human and Animal Physiology
   Wageningen University
   The Netherlands.

4. Laboratory of Bioinformatics
   Wageningen University
   The Netherlands

Submitted for publication in *Molecular Biology and Evolution*

Abstract

Background

The chicken is an important agricultural and avian-model species. Chicken genes are largely annotated based on sequence conservation without further functional evidence. A survey of gene expression in a range of different tissues under normal physiological conditions will support functionality for these genes.

Results

We carried out a gene expression survey in eight major chicken tissues using whole genome microarrays. A global picture of gene expression is presented for the eight tissues and tissue specific as well as common gene expression was identified. A Gene Ontology (GO) term enrichment analysis shows that tissue-specific genes are enriched with GO terms reflecting the physiological functions of the specific tissue and housekeeping genes are enriched with GO terms related to essential biological functions. Comparisons of genomic features between tissue-specific genes and housekeeping genes show that housekeeping genes are more compact. Furthermore, comparisons of gene expression in a panel of five common tissues between chicken, mouse and frog showed that the expression patterns across tissues are conserved for orthologous genes compared to random gene pairs within each pair-wise comparison.

Conclusions

Using whole genome microarrays to survey gene expression across eight normal chicken tissues, we observed tissue-specific patterns of expression for many genes. Commonly expressed genes were more compact, suggesting selection pressure on expression economy. A comparative analysis of gene expression among mouse, chicken, and frog showed evolutionary conservation of the expression patterns of orthologous genes.

Background

The chicken is an important model species for evolutionary and developmental biology, immunology, genetics, as well as for agricultural science. The completion of a draft sequence of the chicken genome [1] represented a landmark in avian genomics and has opened new possibilities to understand gene function and its relationship to physiology. Often gene functions of chicken genes were annotated based on sequence conservation without further functional evidence. A survey of gene expression in a range of different tissues under normal physiological conditions, therefore, would provide additional support for the potential function of many of the chicken genes.

Several studies, using chicken as a model, have compared gene expression differences under different infection treatments using microarrays [2-6]. Most of these studies surveyed gene expression in a single tissue (mostly immune related) and identified genes differentially expressed between two or more conditions (control vs. treatments) in the tissue of interest. However, the identified marker genes for diagnosis and molecular targets for vaccines will depend on knowledge not only of the genes expressed in the diseased tissues of interest, but also on detailed information about the expression of the corresponding genes across different normal tissues. In chicken, the global expression pattern of the genes under normal physiological conditions across a range of tissues and developmental stages needs to be surveyed to provide a global picture of the chicken transcriptome. This information would also provide a baseline for future expression studies on diseases and other traits in chickens. Meanwhile, the global distribution of gene expression among several tissues would help us to identify genes with housekeeping functions and genes with tissue-specific functions. In humans housekeeping genes were found to have relatively shorter introns, untranslated regions and coding sequences, suggesting a selection for compactness of genes that show a wide tissue distribution of expression [7, 8]. We wanted to establish this observation in chicken, and study the mechanism underneath this observation in chicken. Furthermore, clustering of highly expressed genes within specific chromosomal regions has been reported in human [9], mouse [10], chicken (chapter 6 of this thesis), and fruit fly [11]. These regions were termed “RIDGEs” (Regions of Increased Gene Expression). RIDGEs were reported to be associated with higher expression, higher gene density, shorter gene introns, shorter genes, and some other genomic features in chicken (chapter 6 of this thesis). Shorter introns were also reported for highly expressed genes in the human genome [12], and the authors hypothesized that transcription efficiency is enhanced when intron length is shorter. In the current study we present the analysis of the relationship between chromosomal locations and widely expressed genes in chicken.

Evolutionary changes in gene expression account for most phenotypic differences between different species. Studies on conservation of global gene expression patterns between human and apes [13], human and mouse [14] and different other vertebrate species [15] have been reported previously. The results of these studies suggested that the gene expressions within mammals and even within vertebrates are globally conserved. Therefore, it is interesting to compare gene expression in birds with gene expression in mammals and amphibians, which are the two distant neighboring species of birds. Using this comparative approach, we tested whether the conservation of gene expression is correlated with species divergence time. Mouse and frog were chosen to represent mammals and amphibians in this comparison.

In this study, we used the ARK-Genomics *G. gallus* 20K oligonucleotide microarray (GEO [16] platform accession: GPL8861) representing most known and predicted chicken genes to investigate global gene expression patterns among 40 tissue samples representing eight adult tissues (brain, bursa of Fabricius, kidney, liver, lung, small intestine, spleen, and thymus) in chicken (5 biological replicates per tissue type). To summarize, the objectives of this study are to address the following questions: 1) Can we add information to non-annotated sequences, 2) what is the distribution of gene expression in chicken? 3) Do genes with distinct breadth of expression (number of tissues where a gene is expressed) show a correlation with certain genomic characteristics in chicken? 4) Are the expression patterns of orthologous genes conserved between species?

Results

Gene expression distribution in different chicken tissues

Normalized intensities were used as gene expression levels and genes were defined as being expressed only when their expression was higher than 99% quantile value of the expression of all negative control spots across all the arrays in this study (Figure 1a) as described by Zhang et al. [17]. The probe annotations were updated by mapping the probe sequences to the current chicken genome assembly (WASHUC 2, May 2006) using the approach as described by Neerincx et al. [18]. In total, 14,900 probes out of the 20460 probes were mapped uniquely to the chicken assembly, representing 8,908 unique genes (8,792 Ensembl genes [19] and 116 Entrez genes [20]).

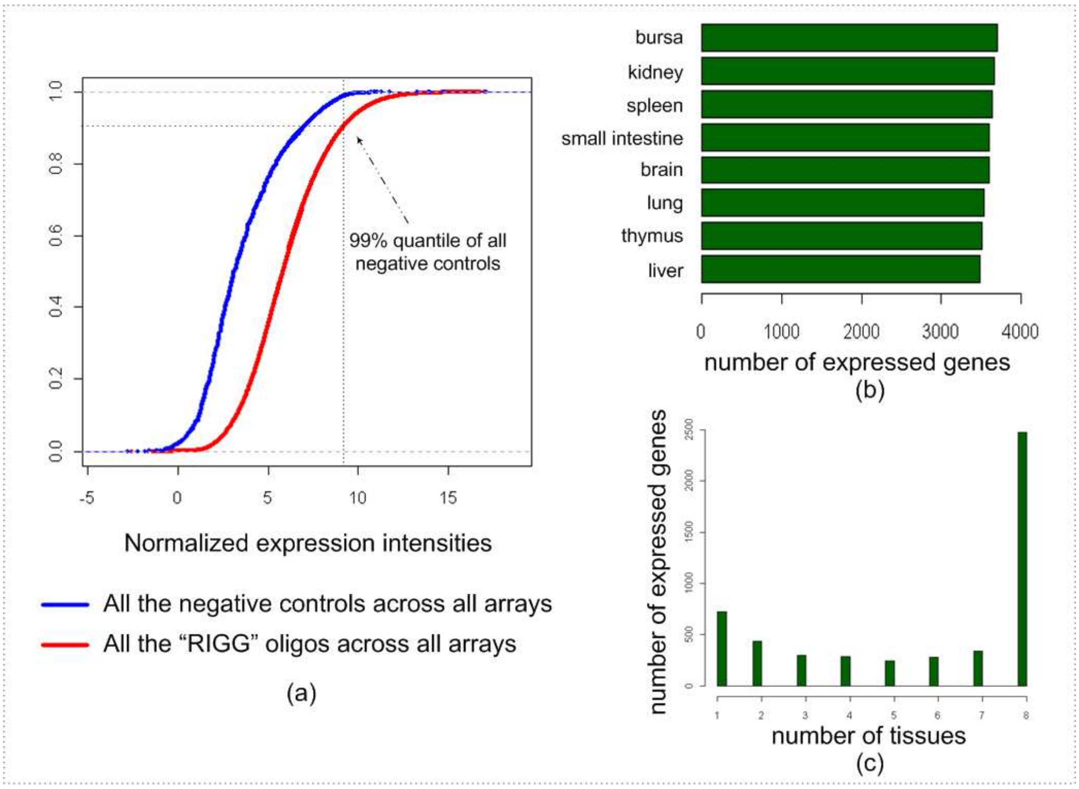


Figure 1. (a). Accumulative plots of arcsihn transformed intensity of genes and negative controls on all the arrays, the red line in Figure 1a indicates all the gene probes on the array and the blue line indicates all the negative control spots across all the arrays. (b). Number of genes expressed in eight chicken tissues (c) Distribution of number of tissues in which genes are expressed (for example, 1 represents the tissue-specific genes, i.e. genes only expressed in one individual tissues, 2 represents that genes are expressed in two tissues out of the eight, and so on.)

The expression data for these genes is available in Additional data file 1. Overall, 57% of the genes are expressed in at least one of the eight tissues (5,086 out of total 8,908 genes represented on the array platform (see materials and methods)). The number of genes expressed in each of the eight individual tissues was similar (Figure 1b) with on average, about 40% of the genes being expressed in each individual tissue type. The distribution of gene expression (number of tissues where a gene is expressed) is shown in Figure 1c. In total, 723 genes showed a single-tissue-specific pattern of expression, whereas 2,476 genes were found to be expressed in all eight tissues (Additional data file 2). In this study, we refer to these 723 genes expressed only in one individual tissue as "tissue-specific genes", and to the 2,476 genes expressed in all eight tissues as "housekeeping genes". The expression levels of housekeeping genes across eight tissues were higher compared to tissue-specific genes (Figure 2).

A Gene Ontology (GO) [21] terms enrichment analysis was performed using GOstats [22] on tissue-specific genes in each tissue type and on the housekeeping genes. The significant (p value < 0.01) GO terms for Biological Process (BP) of the tissue-specific genes are shown in Additional data file 3. The GO terms enriched for each tissue-specific gene list nicely correlates with the physiological function of the individual organs. For example, brain specific genes have enriched GO terms like "neurogenesis", "nervous system development", "neurotransmitter secretion", and "learning" while liver specific genes have enriched GO terms like "blood coagulation", "response to wounding" and "positive regulation of angiogenesis", functions one typically might expect from brain and liver tissues, respectively.

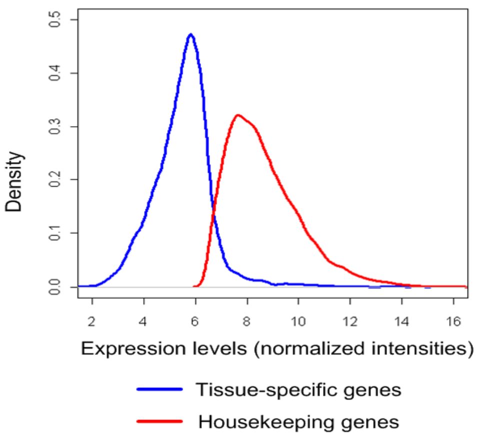


Figure 2. Density plot of expression levels for tissue-specific genes (blue line) and housekeeping genes (red line) across 8 chicken tissues.

The significant (p value<0.01) GO terms (BP) of housekeeping genes indicate that these widely expressed genes are mainly involved in a number of essential biological processes for maintaining a cell (Additional data file 4). GO terms like "translation", "protein folding", "protein localization", "rRNA processing" and "regulation of gene expression" indicate that most of these housekeeping genes are involved in regulation of transcription and translation.

Expression distribution of un-annotated probes

In the above analysis, we only included the probes on the microarray platform which were mapped to Ensembl gene IDs or Entrez gene IDs. There are 5,357 probes, that have a single perfect hit in the current chicken genome assembly but that still lack any annotation even after applying the re-annotation methodology described by Neerincx et al. [18]. About 47.7% (2,556 out of 5,357 probes) of these none-gene probes were expressed in at least one tissue out of the 8 tissues (additional data file 5) with 435 probes being expressed in only one tissue (Additional data file 6), and 1,189 probes being expressed in all 8 tissues (Additional data file

7) . The expression distribution among the eight chicken tissues of these 2,556 expressed un-annotated probes is very similar to the 5,086 expressed annotated genes (Additional data file 8).

Several of the brain specific probes were partly mapped to the last exons of the genes or to the regions directly downstream of the last exon of an annotated gene (see several examples in Additional data file 9). In total, 165 probes were identified to be specifically expressed in brain tissues (Additional data file 9), about 65% of these probes were mapped to cDNA clones/ESTs derived from chicken brain tissues, heads of embryos, and whole embryos of chicken. Probes RIGG12111, RIGG13067, and RIGG11000 from these 165 brain-specific probes show three different typical situations of mapping of these 165 brain-specific probes (Additional data file 10), where probes are either having hits which are partly overlap with exons of known genes or are having hits in genomic regions where no annotation was present previously. For example, probe RIGG10235 (Additional data file 11) was partly mapped to the last exon of Ensembl gene ENSGALG00000000918 (CCDC103), the 1-to-1 human ortholog of CCDC103 known to be expressed in brain tissues. Likewise, probe RIGG16362 (Additional data file 12) was mapped to the region downstream of the last exon of ENSGALG00000011560, whose 1-to-1 human orthologous gene (PACRG) was reported to be a component of the ependymal cilia that may play an important role in motile cilia development and/or function in the central nervous system (CNS) [23].

Housekeeping genes are compact compared to tissue-specific genes

Besides the distinct functions of housekeeping genes compared to tissue-specific genes, we also examined the genomic features, e.g. gene length, coding sequence length, average exon length, average intron length, and intergenic region length, of both the 2,476 housekeeping genes and the 723 tissue-specific genes. Significant differences of gene length (p value=1.4 x 10-13, Wilcoxon Rank Sum Test), coding sequence length (p value=3.1 x 10-13), average intron length (p value=3.7 x 10-13), and intergenic region length (p value=5.8 x 10-9) were found between housekeeping and tissue-specific genes (Figure 3), whereas no differences are observed for the average exon length (p value=0.96) of these two groups of genes. These results suggest that in chicken housekeeping genes are relatively more compact than tissue-specific genes.

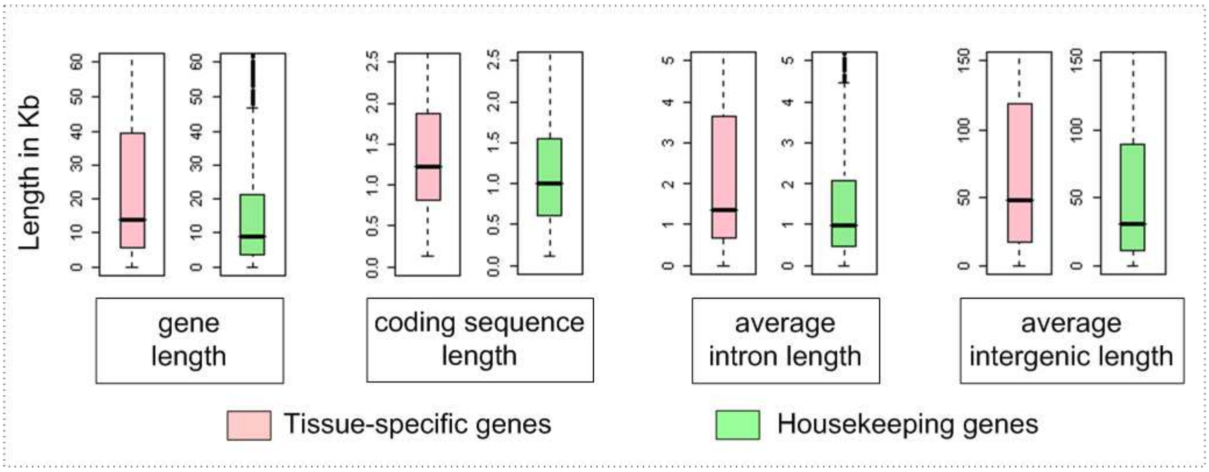


Figure 3. Box plot of gene lengths for tissue-specific genes and housekeeping genes identified based on gene expression in eight chicken tissues.

Chicken housekeeping genes are significantly more located in RIDGEs

A chicken transcriptome map is described previously [24],and regions with clusters of the most highly expressed genes, covering about 10% of the chicken genome, so called “RIDGEs”, are identified. We checked the genomic locations of all 2,476 housekeeping genes in this study and found that about 31% (741 genes) of the housekeeping genes are located within RIDGEs in the chicken genome. To test the significance of the favorable distribution of housekeeping genes within RIDGEs, we performed a random permutation analysis by sampling 2,476 random genes for 1000 times from all 8,908 genes included in this analysis and computed the percentages of random genes being located within RIDGEs. Compared to housekeeping genes, randomly selected genes are much less often located in RIDGEs (13±0.6%, mean±sd). Therefore, the genomic locations of house-keeping genes show a higher overlap with RIDGEs across the chicken genome.

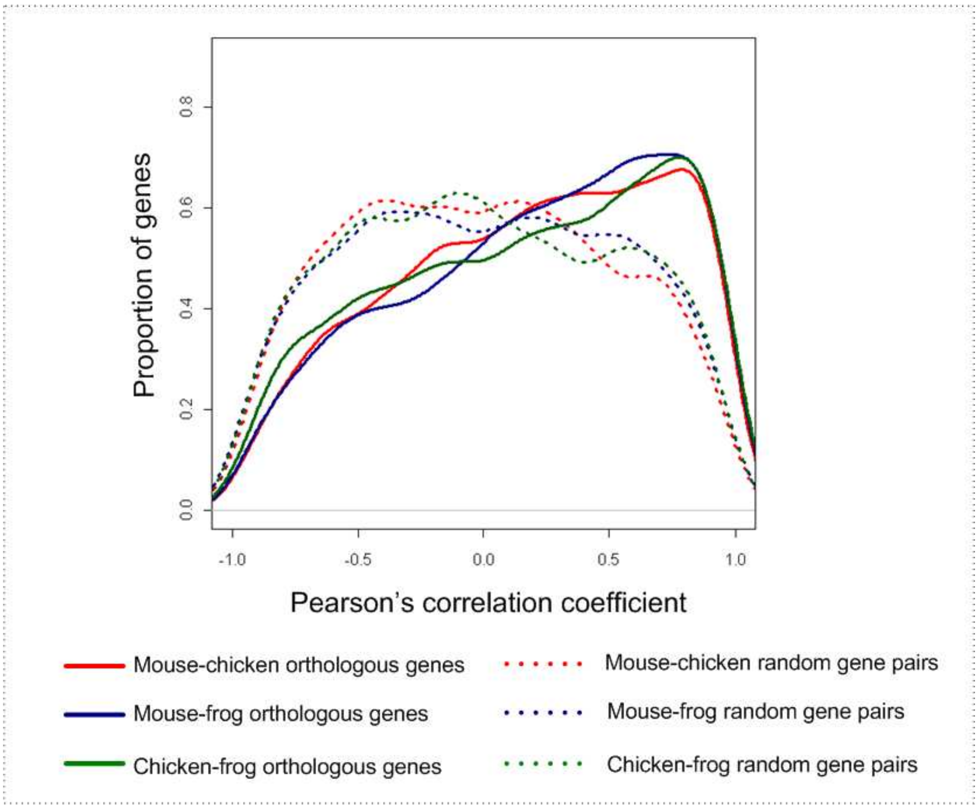


Figure 4. Distribution of gene expression correlation coefficients of orthologous gene pairs and random gene pairs in pair-wise comparisons among mouse, chicken, and frog.

Expression of orthologous genes is conserved in vertebrates

Conservation of gene expression was compared by checking the 3,892 1:1:1 orthologous genes in mouse, chicken and frog. Pair-wise comparisons were performed among the three species and significant conservation of gene expression was found when comparing orthologous gene pairs to random gene pairs within each pair-wise comparison (Figure 4). When, within each comparison, the correlation between the gene expressions of an orthologous gene pair was higher than 95% quantile of random gene pairs (as background), we labeled the orthologous gene pair as having a conserved expression pattern. In total, 11.3% (439 genes out of 3,892 genes) chicken-mouse orthologous genes, 10.9% (425 genes) chicken-frog orthologous genes, and 5.01% (195 genes) mouse-frog orthologous genes show a conserved gene expression profile within each pair-wise comparison.

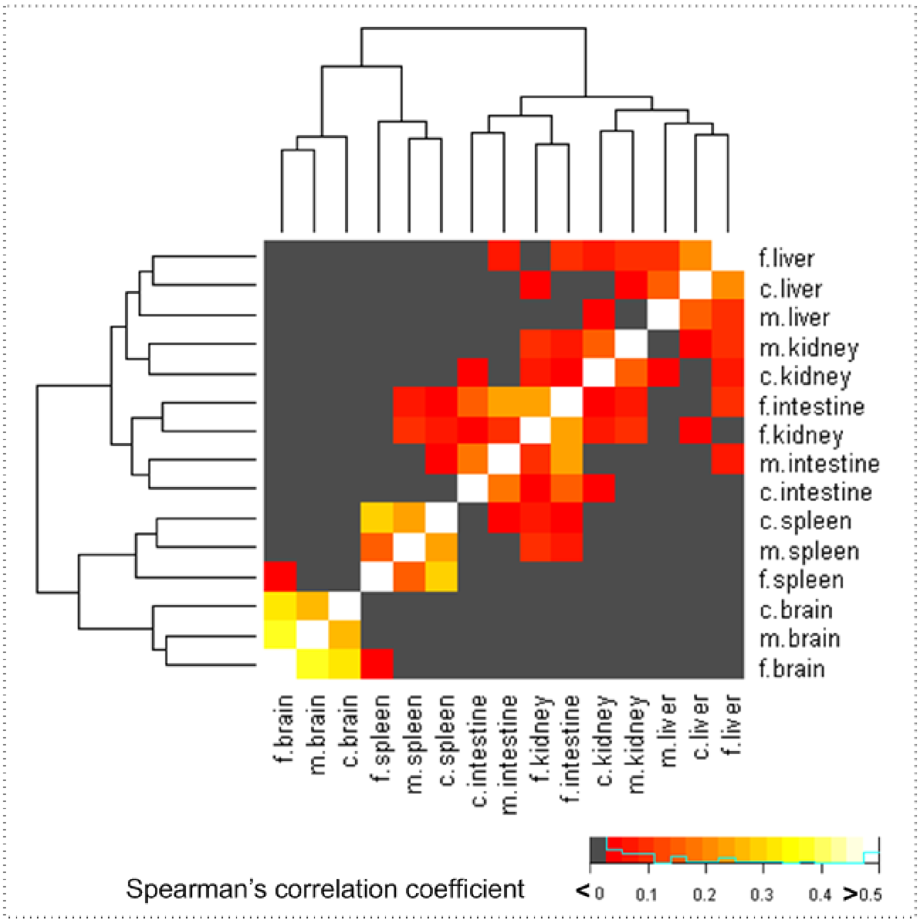


Figure 5. Heat map of correlation coefficients (Spearman) between five common tissues (m: mouse, c: chicken, and f: frog) in three different species.

Homologous tissues are more similar in vertebrates in terms of expression

Besides testing conservation of gene expression of orthologous genes between species, we also tested whether homologous tissues (for example, brain tissues in mouse, chicken, and frog) are more similar to each other compared to non-homologous tissues. After transforming gene expression intensities to relative expression ratios (RA) across the same panel of tissues, a comparison between global gene expression profiles among tissues in different species was possible. The rank correlation coefficient among different tissues showed that homologous tissues in three different species are more similar compared to non-homologous tissues (Figure 5); especially brain tissues are highly correlated within the three species indicating evolutionary constraints are posed on brain gene expression profiles. In contrast, kidney showed a relatively low conservation.

Discussion

Gene expression distribution in various chicken tissues

The main objective of this study was to survey gene expression profiles across a set of eight normal chicken tissues. We present a microarray expression dataset surveying about 8,792 chicken Ensembl genes across 8 different chicken tissue types in 5-fold (brain, bursa of Fabricius, kidney, liver, lung, small intestine, spleen, and thymus). For most genes the distribution of expression is observed across several different tissues (Figure 1c). For 723 genes, a single-tissue-specific pattern is seen, while 2,476 genes were found to be expressed in all eight tissues. The genes with expression across the eight tissues indicate their universal biological function in cells and therefore can be considered as genes with "housekeeping functions", although a proper definition of such genes would require a comprehensive sampling of tissues for the whole organism. The GO term enrichment analysis of housekeeping genes show the enriched biological processes GO terms like "translation", "protein folding", "protein localization", "rRNA processing" and "regulation of gene expression" (Additional data file 4). This confirmed that our definition of "housekeeping gene" was vald.

Potential shortcomings of the current gene models in the chicken genome

The wide distribution of expression for the un-annotated 2,556 probes among the eight tissues implies that many genes/transcripts in the chicken genome are not well annotated in this genome assembly. These 2,556 probes can be used as expressed evidence for potential gene/transcript prediction in the genomic regions where they were uniquely mapped in the chicken genome. The 65% identified brain-specific probes were designed based on cDNA clones/ESTs derived from chicken brain tissues, heads of embryos, and whole embryos of chicken, suggesting that there are still many transcribed regions in the chicken genome that have not yet been annotated in the current gene models. Two examples of probe RIGG10235 and RIGG16362 suggest that the current prediction of 3' UTR of chicken genes is more difficult, i.e. the 3' UTR of chicken genes are not very accurately predicted in the current assembly, and all the other expressed probes not mapped to known genes imply that the chicken genome contains a large number of still un-annotated transcribed regions.

Housekeeping genes are compact compared to tissue-specific genes

The on average smaller size observed for the housekeeping genes is due to both a shorter coding sequence as well as a shorter intron length. Furthermore, the smaller size of the intergenic region also contributes to a higher gene density of the areas containing the housekeeping genes, suggesting a selection for compactness, which has also been reported in human [7, 8], this might reduces the costs of transcription of housekeeping genes. It has been shown that translation is more costly than transcription [25], and the shorter length of the coding sequences in housekeeping genes is likely the result of selection for economy of translation. On the other hand, the tissue-specific genes are longer, because of their higher number of functional domains and relative more complex protein architecture as was previously reported in human [8]. Likewise, regulation of expression of these genes in a number of specific tissues might have resulted in a large number of cis-regulatory elements and would need larger regulatory “spaces” resulting in larger introns and intergenic regions.

Housekeeping genes are in favor of being located in RIDGEs in the chicken genome

The hypothesis for the existence of RIDGEs is that evolution favors highly expressed genes to be co-localized, as transcription of one gene would help the chromatin of neighboring genes to “open up” during transcription. The favorable distribution of housekeeping genes within RIDGEs again indicates that these genes need to be expressed at relative higher levels (Figure 2) and at a larger number of physiological conditions (“housekeeping functions”)

Expressions profiles of orthologous genes are conserved in vertebrates

In contrast to direct sequence comparisons of orthologous genes, the comparison of the gene expression profiles of orthologous genes has a number of caveats. First of all, the expression levels of genes are dynamic and change with developmental and physiological state. Secondly, the tissue samples collected in this study, as well as those in the other two published gene expression surveys used in this study are only a part of all organs, representing the average of millions of cells of several different types.

Nevertheless, the expression of orthologous genes is generally well conserved as compared to random gene pairs (Figure 4). If gene expression were to evolve in accordance with neutral theory [26], the expression of orthologous genes would be the same as random gene pairs, while our results suggest that gene expression is under some selection constraint during evolution. The overall correlation distributions of orthologous gene expressions are similar when comparing each pairs among the three species mouse, frog and chicken.

Conclusions

We have used whole genome microarrays to survey gene expression across eight normal chicken tissues. Most genes show tissue-specific patterns of expression and do not show any clear preference for being clustered in specific regions of the genome. Housekeeping genes on the other hand are more likely to co-localize with other abundantly or highly expressed genes. There seems to be selection pressure on economy in genes with a wide tissue distribution (housekeeping genes), i.e. these genes are more compact. A comparative analysis of gene expression among mouse, chicken, and frog showed that the expression patterns of orthologous genes are conserved between mammals, birds, and amphibians during evolution.

Materials and Methods

Tissue sample preparation

In total, 5 healthy ten week old chickens were used for this study. The animal experiment was approved by The Institutional Animal Care and Use Committee of Wageningen University. All tissue samples (brain, bursa of Fabricius, kidney, liver, lung, small intestine, spleen, and thymus) were collected and immediately put into the RNA Stabilization Reagent RNAlater (Qiagen, Valencia, CA, USA), followed by incubation at +4℃ overnight, then storage at -80℃ until use.

RNA isolation, labeling and hybridizations

Total RNA was isolated using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions, followed by a subsequent sample "clean-up" using RNeasy Mini Kit (Qiagen). RNA quantity was measured using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, USA). The quality and integrity of the RNA was analyzed using the Agilent Bioanalyzer 2100 (Palo Alto, CA, USA), RNA was amplified using MessageAmp™ II aRNA Kit (Ambion, Foster City, CA, USA) and cRNA was further used for chemical coupling with ULS-Cy3/Cy5 (ULS™ aRNA labeling kit; Kreatech, Amsterdam, Netherlands). After coupling and purification the cRNA concentration and fluorescent incorporation was quantified using the Nanodrop Spectrophotometer. One µg of each labeled cRNA sample was used to hybridize on the Ark-Genomics *G.gallus* 20k array. The hybridizations were done overnight on a GeneTAC hybridization station (Genomic Solutions, Holliston, MA, USA). Hybridized arrays were scanned using Agilent DNA microarray scanner (Agilent, Santa Clara CA, USA). A common reference design was used in this study. The common reference was made by pooling total RNA samples from all individual samples, and each individual sample was hybridized against the common reference on the same array slide. In all 40 arrays (5 biological replicates per tissue type), cRNA of individual tissue samples was always labeled using Cy3 (green), and the cRNA of the common reference samples were always labeled using Cy5 (red).

Array probe re-annotation

The ARK-Genomics G. gallus 20K array platform, used in this study, contains 20,460 unique oligonucleotide probes (GEO accession GPL8861, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=tjwjpscyceqawjk&acc=GPL8861). The probe sequences were mapped to the current chicken genome assembly (Ensembl Genome database release 50, WASHU2 assembly, May 2006) using the method described previously by Neerincx et al. [18]. In total, 14,900 probes were mapped to a unique position in the current assembly, corresponding to 8,792 unique ensembl genes in the Ensembl Genome Database and 5,357 probes were mapped to unique positions in the genome without a link to an ensembl gene.

Microarray data processing, normalization, and statistical analysis

Scanned TIFF images were analyzed using GenePix 6.0 (Axon, Sunnyvale, CA, USA), and results were saved as GenePix Result (*.gpr) files. We used R/Bioconductor package Limma to analyze the array data. The *.gpr files were imported into R (version 2.8.0), median values of both foreground and background intensities were extracted and used in the analysis. We gave any spot with FLAG-value less than -50 (these spots were flagged as “bad spot” by GenePix program or manually) a weight of 0.01, and all the other spots we gave weights of 1. The raw data was normalized in R using variance stabilizing normalization (VSN) methods implemented in package vsn [27]. The normalized intensities of the green channel (representing all individual tissue samples) were used as gene expression data in the analysis and the data points for those spots (both genes and negative controls) with low weight (0.01) were removed in further analysis. The gene expression data was first averaged within each tissue type among the five biological replicates, and then the gene expression data for probes targeting the same Ensembl genes/entrez gene were averaged.

Gene Ontology term enrichment analysis

All the genes having a chicken Ensembl gene ID were mapped to their 1-to-1 human orthologous genes using Bioconductor package biomaRt [28] through the Ensembl Genome Database. The GO term enrichment analysis was subsequently performed using human gene annotation using R package GOstats [22]. A conditional hypergeometric test algorithm provided within GOstats package was applied to GO enrichment analysis. The conditional hypergeometric test identifies a GO term as significant if there is evidence beyond that provided by its significant children. Only the enriched GOBP terms with raw p-values < 0.01 were used for biological interpretation in this study.

Comparing 1-1-1 orthologous gene expression conservation

Orthologous genes for mouse (*Mus musculus*), chicken (*Gallus gallus*), and frog (*Xenopus tropicalis*) were downloaded from Ensembl. The normalized gene expression data for mouse and frog were downloaded from the functional landscape of mouse gene expression website [29] and the Conservation of Core Gene Expression in Vertebrate Tissues: Supplementary Data website [30], respectively. The expression data of chicken in this study was normalized using the same method as used in these two previous studies [15, 17]. The gene expression data from different species using different species-specific microarray platforms are not directly comparable., To enable cross-species gene expression comparisons, we used relative mRNA abundance among tissues (RA) introduced by Liao and Zhang [14]. Gene expression levels were calculated as ratios between the expression intensity of gene X in one particular tissue divided by sum of expression intensities of gene X in all tissues included in the analysis.

Abbreviations

GO: Gene Ontology
GOBP: Gene Ontology Biological Process
RIDGE: Regions of Increased Gene Expression
RA: Relative mRNA abundance
CNS: Central Nervous System

Authors’ contributions

HN, RPMAC, and AL conducted the experiment. HN performed the data analysis and drafted the manuscript. HJM, PBTN, JAML, EMS and JK helped with data analysis, MAMG supervised the project, and all authors were involved in improving the manuscript. The final version of the manuscript was approved by all the authors.

Acknowledgements

This work was financed by European Union grant FOOD-CT-2004—506416 (EADGENE). The authors would like to thank Basav Hangalapura from Adaptation Physiology group of Wageningen University for helping collecting chicken tissues, we also thank Annelies Bunschoten from Human and Animal Physiology of Wageningen University for helping with RNA quality check in the lab.

References

1. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.
2. van Hemert S, Hoekman AJ, Smits MA, Rebel JM: **Gene expression responses to a Salmonella infection in the chicken intestine differ between lines**. *Vet Immunol Immunopathol* 2006, **114**(3-4):247-258.
3. van Hemert S, Hoekman AJ, Smits MA, Rebel JM: **Immunological and gene expression responses to a Salmonella infection in the chicken intestine**. *Vet Res 2007*, **38**(1):51-63.
4. Kim DK, Hong YH, Park DW, Lamont SJ, Lillehoj HS: **Differential immune-related gene expression in two genetically disparate chicken lines during infection by Eimeria maxima**. *Dev Biol* (Basel) 2008, **132**:131-140.
5. Kim DK, Lillehoj HS, Hong YH, Park DW, Lamont SJ, Han JY, Lillehoj EP: **Immune-related gene expression in two B-complex disparate genetically inbred Fayoumi chicken lines following Eimeria maxima infection**. *Poult Sci* 2008, **87**(3):433-443.
6. Morgan RW, Sofer L, Anderson AS, Bernberg EL, Cui J, Burnside J: **Induction of host gene expression following infection of chicken embryo fibroblasts with oncogenic Marek's disease virus**. *J Virol* 2001, **75**(1):533-539.
7. Eisenberg E, Levanon EY: **Human housekeeping genes are compact**. *Trends Genet* 2003, **19**(7):362-365.
8. Vinogradov AE: **Compactness of human housekeeping genes: selection for economy or genomic design?** *Trends Genet* 2004, **20**(5):248-253.
9. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA et al: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains**. *Science* 2001, **291**(5507):1289-1292.
10. Mijalski T, Harder A, Halder T, Kersten M, Horsch M, Strom TM, Liebscher HV,

Lottspeich F, de Angelis MH, Beckers J: **Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues**. *Proc Natl Acad Sci U S A* 2005, **102**(24):8621-8626.

11. Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI: **Large clusters of co-expressed genes in the Drosophila genome**. *Nature* 2002, **420**(6916):666-669.
12. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: **Selection for short introns in highly expressed genes**. *Nat Genet* 2002, **31**(4):415-418.
13. Khaitovich P, Enard W, Lachmann M, Paabo S: **Evolution of primate gene expression**. *Nature reviews* 2006, **7**(9):693-702.
14. Liao BY, Zhang J: **Evolutionary conservation of expression profiles between human and mouse orthologous genes**. *Molecular biology and evolution* 2`006, 23(3):530-540.
15. Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M et al: **Conservation of core gene expression in vertebrate tissues**. *Journal of biology* 2009, **8**(3):33.
16. **National Center for Biotechnology Information (NCBI) Gene Expression Omnibus**: [http://www.ncbi.nlm.nih.gov/geo/]
17. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E et al: **The functional landscape of mouse gene expression**. *Journal of biology* 2004, **3**(5):21.
18. Neerincx PB, Rauwerda H, Nie H, Groenen MA, Breit TM, Leunissen JA: **OligoRAP - an Oligo Re-Annotation Pipeline to improve annotation and estimate target specificity**. *BMC Proc* 2009, **3 Suppl 4**:S4.
19. **Ensembl Genome Database**: [http://www.ensembl.org/]
20. **Entrez Gene**: [http://www.ncbi.nlm.nih.gov/gene]
21. **Gene Ontology website**: [www.geneontology.org]
22. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association**. *Bioinformatics* 2007, **23**(2):257-258.
23. Wilson GR, Tan JT, Brody KM, Taylor JM, Delatycki MB, Lockhart PJ: **Expression and localization of the Parkin co-regulated gene in mouse CNS suggests a role in ependymal cilia function**. *Neurosci Lett* 2009, **460**(1):97-101.
24. Nie H, Crooijmans RP, Bastiaansen JW, Megens HJ, and Groenen MA: **Regional regulation of transcription in the chicken genome**. *BMC Genomics* 2010, **11**:28.
25. Hulbert AJ, Else PL: **Mechanisms underlying the cost of living in animals**. *Annu Rev Physiol* 2000, **62**:207-235.
26. Kimura M: **Evolutionary rate at the molecular level**. *Nature* 1968, **217**(5129):624-626.

27. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**. *Bioinformatics* 2002, **18 Suppl 1**:S96-104.

28. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis**. *Bioinformatics* 2005, **21**(16):3439-3440.

29. **The functional landscape of mouse gene expression**: [http://hugheslab.med.utoronto.ca/Zhang]

30. **Conservation of Core Gene Expression in Vertebrate Tissues: Supplementary Data**: [http://hugheslab.ccbr.utoronto.ca/supplementarydata/vertebrate_expression]

Additional data files

Additional data file 1: expression data of 8908 genes in eight chicken tissues
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%201.xls

Additional data file 2: Lists of tissue specific genes and housekeeping genes
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%202.xls

Additional data file 3: GO enrichment results of tissue-specific genes
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%203.xls

Additional data file 4: GO enrichment result of housekeeping genes
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%204.xls

Additional data file 5: 2566 expressed probes not annotated to known genes
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%205.xls

Additional data file 6: probes not mapped to known genes with tissue specific expression patterns
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%206.xls

Additional data file 7: probes not mapped to known genes with housekeeping expression patterns
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%207.xls

Additional data file 8: expression distribution of probes not mapped to annotated genes
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%208.tif

Additional data file 9: 165 brain-specific probes not mapped to known genes
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%209.xls

Additional data file 10: few examples of none-gene probes
http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%2010.tif

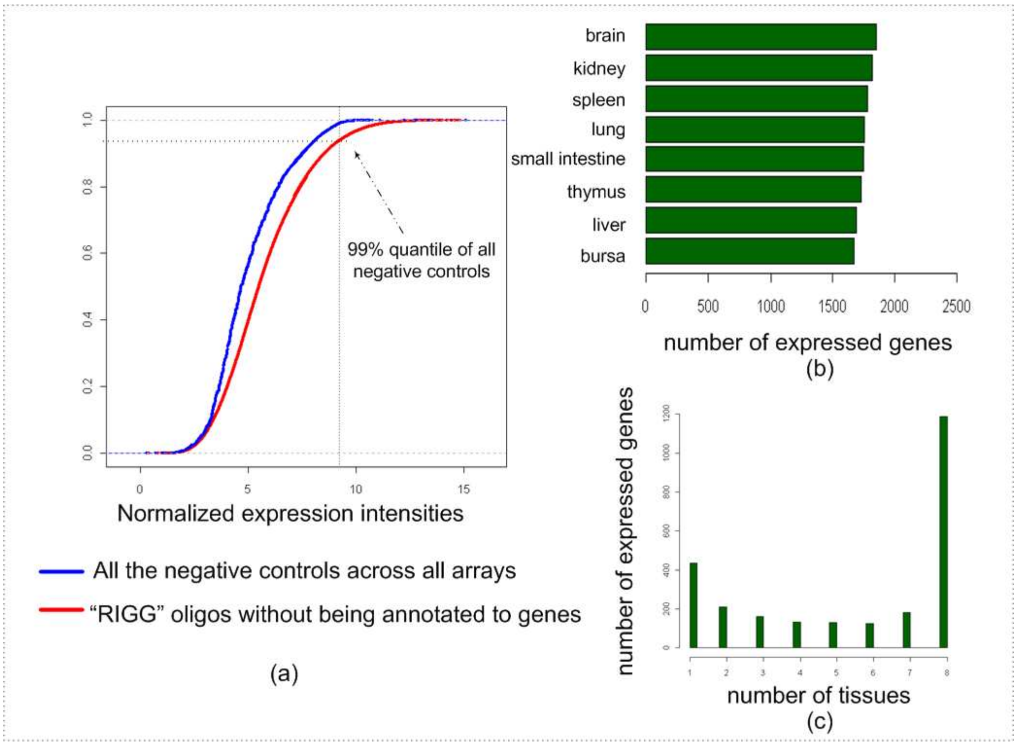
Additional data file 11: probe RIGG10235

http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%2011.tif

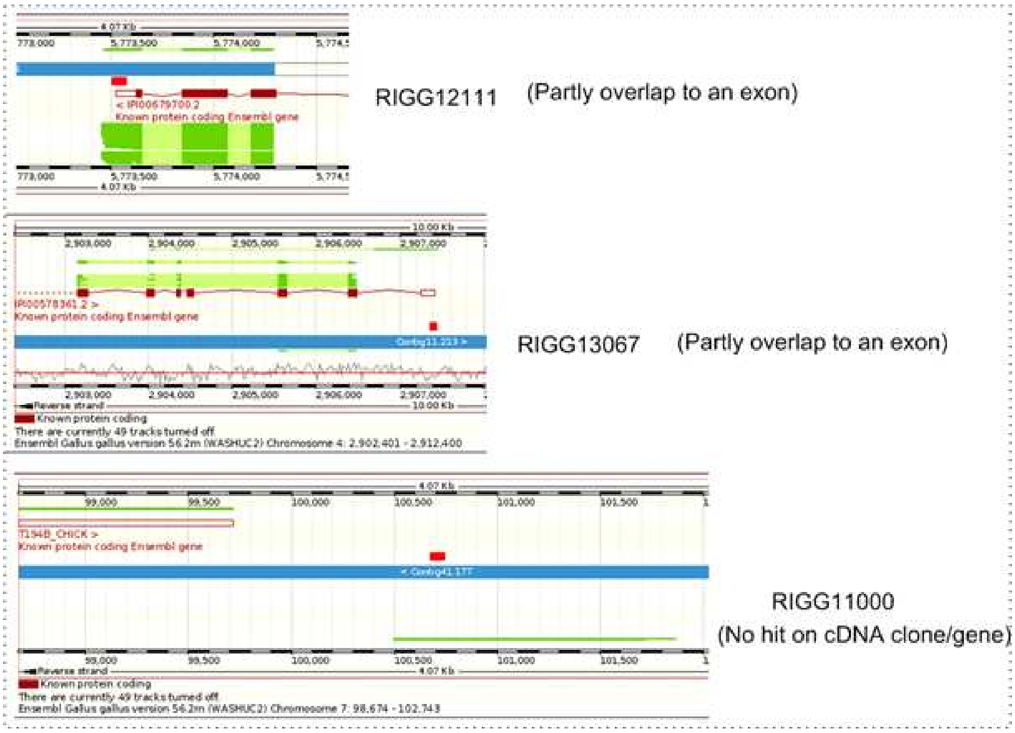
Additional data file 12: probe RIGG16362

http://abgc.asg.wur.nl/nie/Chapter_4/Additional%20data%20file%2012.tif

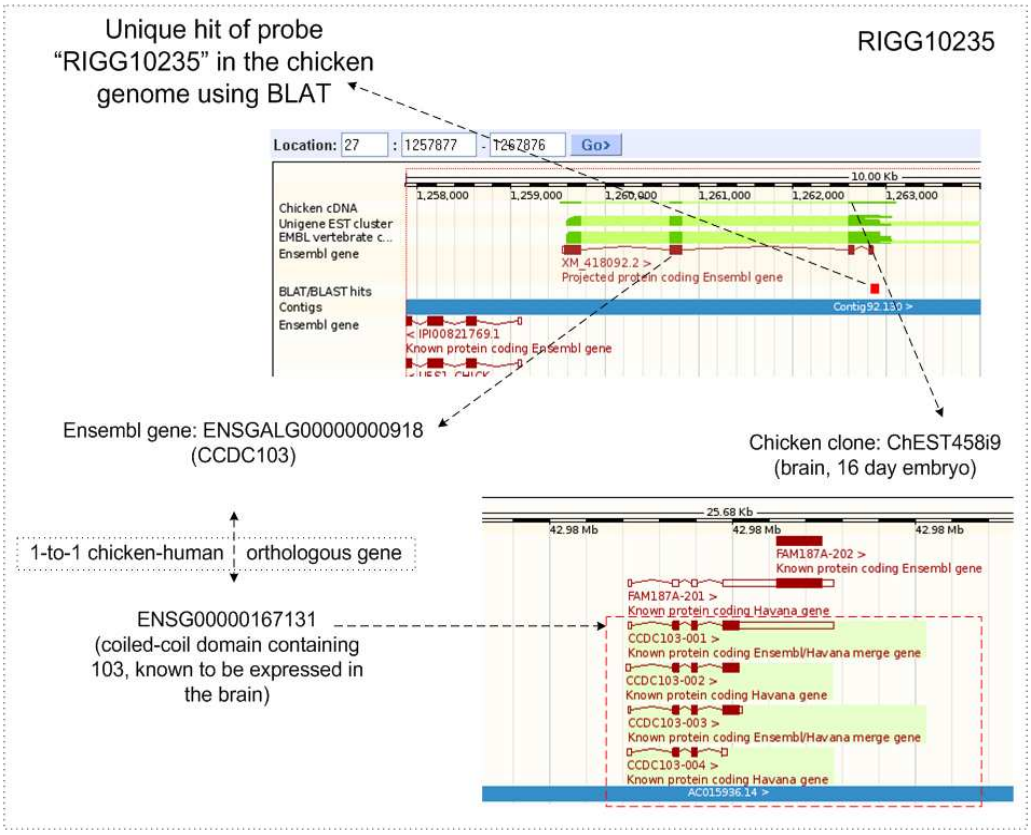
Additional data file 5. Expression distribution of probes not mapped to annotated genes



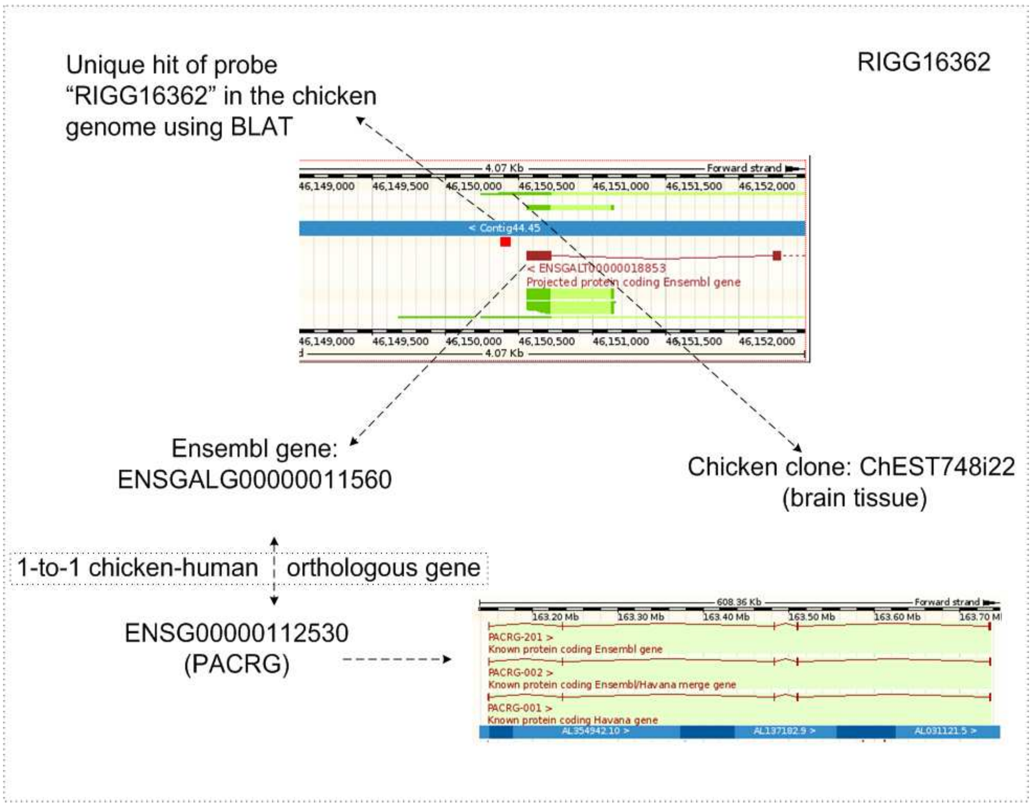
Additional data file 10. Few examples of none-gene probes


4.07 Kb
5,773,500
5,774,000
< IPI00679700.2
Known protein coding Ensembl gene
RIGG12111
(Partly overlap to an exon)
10.00 Kb
2,903,000
2,904,000
2,905,000
2,906,000
2,907,000
IPI00578361.2 >
Known protein coding Ensembl gene
RIGG13067
(Partly overlap to an exon)
Reverse strand
Known protein coding
There are currently 49 tracks turned off
Ensembl Gallus gallus version 56.2m (WASHUC2) Chromosome 4: 2,902,401 - 2,912,400
4.07 Kb
99,000
99,500
100,000
100,500
101,000
101,500
T194B_CHICK >
Known protein coding Ensembl gene
RIGG11000
(No hit on cDNA clone/gene)
Reverse strand
Known protein coding
There are currently 49 tracks turned off
Ensembl Gallus gallus version 56.2m (WASHUC2) Chromosome 7: 98,674 - 102,743

Additional data file 11. Probe RIGG10235


Unique hit of probe "RIGG10235" in the chicken genome using BLAT
RIGG10235
Location: 27 : 1257877 - 1267876 Go>
10.00 Kb
1,258,000
1,259,000
1,260,000
1,261,000
1,262,000
1,263,000
Chicken cDNA
Unigene EST cluster
EMBL vertebrate c...
Ensembl gene
XM_418092.2 >
Projected protein coding Ensembl gene
BLAT/BLAST hits
Contigs
Ensembl gene
Contig92.130 >
< IPI00821769.1
Known protein coding Ensembl gene
Ensembl gene: ENSGALG00000000918 (CCDC103)
Chicken clone: ChEST458i9 (brain, 16 day embryo)
1-to-1 chicken-human orthologous gene
ENSG00000167131 (coiled-coil domain containing 103, known to be expressed in the brain)
25.68 Kb
42.98 Mb
42.98 Mb
42.98 Mb
FAM187A-202 >
Known protein coding Ensembl gene
FAM187A-201 >
Known protein coding Havana gene
CCDC103-001 >
Known protein coding Ensembl/Havana merge gene
CCDC103-002 >
Known protein coding Havana gene
CCDC103-003 >
Known protein coding Ensembl/Havana merge gene
CCDC103-004 >
Known protein coding Havana gene
AC015936.14 >

Additional data file 12. Probe RIGG16362


Unique hit of probe "RIGG16362" in the chicken genome using BLAT
RIGG16362
4.07 Kb
Forward strand
46,149,000
46,149,500
46,150,000
46,150,500
46,151,000
46,151,500
46,152,000
< Contig44.45
< ENSGALT00000018853
Projected protein coding Ensembl gene
Ensembl gene: ENSGALG00000011560
Chicken clone: ChEST748i22 (brain tissue)
1-to-1 chicken-human orthologous gene
ENSG00000112530 (PACRG)
608.36 Kb
163.20 Mb
163.30 Mb
163.40 Mb
163.50 Mb
163.60 Mb
163.70 Mb
PACRG-201 >
Known protein coding Ensembl gene
PACRG-002 >
Known protein coding Ensembl/Havana merge gene
PACRG-001 >
Known protein coding Havana gene
AL354942.10 >
AL137182.9 >
AL031121.5 >

Chapter 5

A Genome-wide Gene Expression Survey in chicken embryos and embryonic tissues

Haisheng Nie[1], Richard P.M.A Crooijmans[1], Ilse .M.L. de Jong[2], Merijn A. G. de Bakker[2], Michael K. Richardson [2], and Martien A.M. Groenen[1]

1. Animal Breeding and Genomics Centre
   Wageningen University
   The Netherlands

2. Institutes of Biology
   Leiden University
   The Netherlands

Manuscript in preparation

Abstract

Background

The chicken embryo has been a popular model in embryology and developmental biology. Despite this fact there is very limited information available about large scale gene expression surveys in different chicken embryonic stages and embryonic tissues to study the molecular mechanism of embryonic development and/or organ differentiation in embryos.

Results

A gene expression survey was conducted using a whole genome chicken 20K oligonucleotide microarray to study the overall gene expression pattern in whole chicken embryos at four different developmental stages (HH stage 3, 10, 15, 22) and in eight different embryonic tissues (brain, bursa of Fabricius, heart, kidney, liver, lung, small intestine, spleen from HH stage 36 embryos) . Developmental stage-specific and tissue-specific genes were identified. A GO enrichment analysis shows that tissue-specific genes correspond to the physiological functions of the tissues. Furthermore, genomic features of genes widely expressed under these 12 conditions confirmed earlier findings (Chapter 4) that widely expressed genes are more compact than tissue-specific genes. A detailed analysis of differentially expressed gene in each pair-wise comparison among different developmental stages also showed gradual changes on gene expression during embryogenesis. Comparisons were performed between tissue-specific genes identified in adult tissues in Chapter 4 and tissue-specific genes identified in embryonic tissues identified in this study. Similarities and differences about organ functions at different developmental stages (adult vs. embryonic stages) are discussed in this study.

Conclusions

In this study, stage- and tissue-specific genes among a variety of embryonic stages and embryonic tissues have been identified. Biological processes at the molecular level were discovered during embryonic developments. Comparisons of functions between organs, on the transcriptomic level, reveal similarities and differences of adult organs and embryonic organs in chicken.

Background

The chick embryo has been a popular model in embryology and developmental biology. The extensive use of the chicken as one of the primary models for developmental biology is due to the easy access of the embryo because development occurs *in ovo* rather than *in utero*, which allows easy manipulation of the incubated eggs and the developing embryo. However, there is very limited information available for genome-wide gene expression profiles in different chicken embryonic stages and embryonic tissues. The completion of a draft sequence of the chicken genome [1] made it possible to develop genome-wide gene expression microarray platforms [2, 3] to survey the expression profiles across different developmental stages/tissues.

The chicken embryonic development process was divided into stages by Hamburger and Hamilton in 1951 [4]. The morphological characteristics are gradually changing during the embryonic development at different HH stages. During embryonic development, various cellular and molecular changes take place under transcriptional regulation. To identify gene expression profiles and search for new candidate genes involved in this developmental process, chicken embryonic gene expression was analyzed with a chicken whole genome 20k oligo-array [3].

To study the molecular mechanisms of chicken embryonic development, we selected chicken whole embryos at HH stage 3+, HH stage 10, HH stage 15, and, HH stage 22 to survey the genome-wide gene expression across these stages. HH stages 3+, HH stage 10, and HH stage 22 represent the three landmark developmental points of embryonic development: gastrulation, limb-bud, and tail-bud respectively. Furthermore, the majority of the organ systems have been established between the sixth day and hatching, much of development is concerned largely with increase in size of existing organs [5]. To investigate transcriptomic differences in different embryonic organs, we also surveyed gene expression across eight major embryonic tissues from HH stage 36, after 10 days incubation.

In this study, we report a larger scale microarray-based survey of gene expression across 4 different chicken embryo stages and 8 different embryonic tissues.

Results

Gene expression distribution in embryo stages and embryonic tissues

Normalized intensities were used as gene expression levels and genes were defined as being expressed only when their expression was higher than the 99% quantile value of the expression of all negative control spots across all the arrays in this study (Figure 1a), the same approach as described in Chapter 4. The probe annotations were updated according to Neerincx et al. [6] as described in Chapter 4,. In total, 14,900 probes out of 20,460 were mapped uniquely to the chicken assembly, representing 8,908 unique genes (8,792 Ensembl genes [7] and 116 Entrez genes [8]) The expression data of these 8,908 genes is available in Additional data file 1. In total, 73% of the genes are expressed in at least one of the 12 tissues (5,086 out of total 8,908 genes represented on the array platform (see materials and methods)). The distribution of gene expression (number of tissues where a gene is expressed) is shown in Figure 1b. In total, 685 genes showed a single-tissue-specific pattern, whereas 3,228 genes were found to be expressed in all 12 tissues (data available in Additional data file 2). In this study, we refer to these 685 genes expressed only in one individual tissue as “stage/tissue-specific genes”, and to the 3,228 genes expressed in all 12 stages/tissues as “housekeeping genes”. The Gene Ontology (GO) [9] terms enrichment analysis was performed using GOstats [10] on tissue-specific genes in each tissue types and on the housekeeping genes. The significant (p value < 0.01) GO terms for Biological Process (BP) of the tissue-specific genes is shown in Additional data file 3. The GO terms enriched for each tissue-specific gene list nicely correlates with the physiological function of the individual organs. For example, HH stage 3 embryos has term “glucocorticoid receptor signaling”, embryonic brain was enriched with terms like “synaptic transmission” and “visual learning”, embryonic bursa was enriched with “activation-induced cell death of T cells” and “inflammatory cell apoptosis”, and spleen was enriched with “immune response”. The significant (p value<0.01) GO terms (BP) of housekeeping genes indicate that these widely expressed genes are mainly involved in a number of essential biological processes for maintaining a cell (Additional data file 4). GO terms like “cell cycle process”, “RNA processing”, “translation”, and “protein folding” indicate that most of these housekeeping across the 12 stages/tissues are involved in cell division and proliferation, this implies that during embryonic development, cell division is the most prominent biological process in embryo/embryonic tissues.

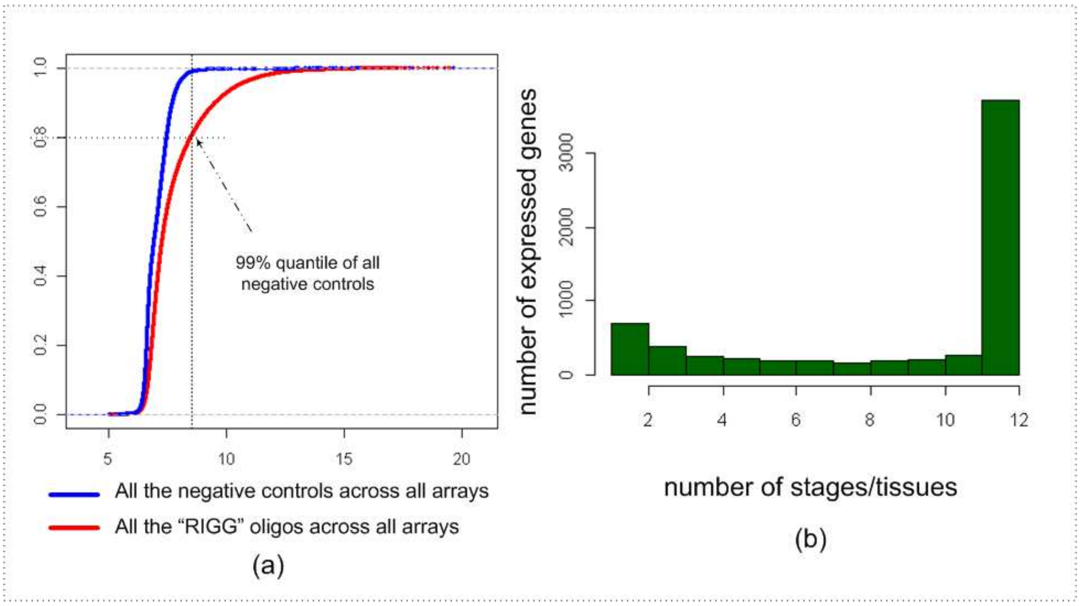


Figure 1. Accumulative plots of normalized intensities of genes and negative controls across all the arrays (c) Distribution of number of embryonic stages/tissues in which genes are expressed.

Characteristics of widely expressed genes

Similar to the analysis that was introduced in Chapter 4, we also examined the genomic features, e.g. gene length, coding sequence length, average exon length, average intron length, and length of the intergenic region, for all 3,228 housekeeping genes and 685 stage/tissue-specific genes. Significant differences of gene length (p value=1.0 x 10-10, Wilcoxon Rank Sum Test), coding sequence length (p value=2.2 x 10-16, Wilcoxon Rank Sum Test), average intron length (p value=3.4 x 10-7, Wilcoxon Rank Sum Test), and length of the intergenic region (p value=0.0004, Wilcoxon Rank Sum Test) were found between housekeeping and tissue-specific genes (Figure 2), whereas no differences are observed for the average exon length (p value=0.96, Wilcoxon Rank Sum Test) of these two groups of genes. These results suggest that in chicken housekeeping genes are relatively more compact than tissue-specific genes.

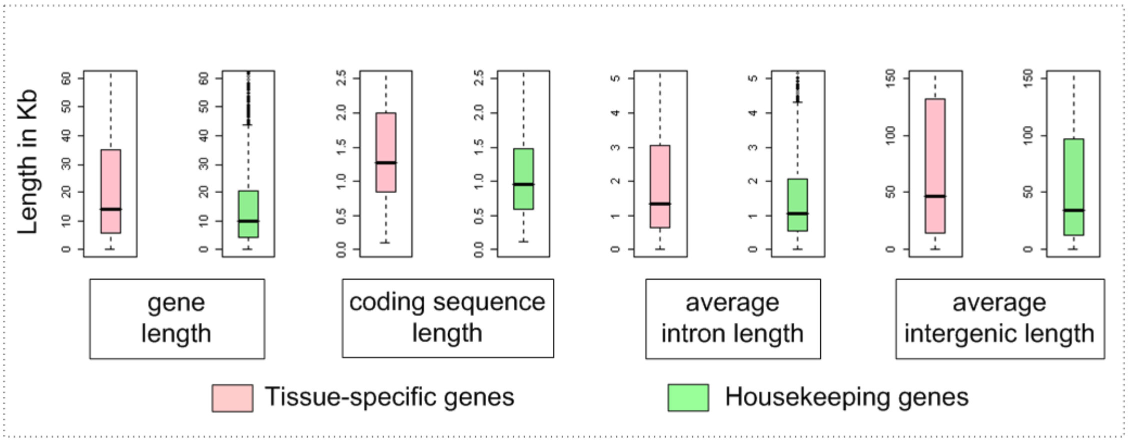


Figure 2. Box plot of several genomic features for tissue-specific genes and housekeeping genes identified based on gene expression in the 12 embryonic stages/tissues.

Identification of biological processes during embryonic development

Embryo development is a continuous process. We decided to test which genes are differentially expressed between two consecutive embryonic stages. We therefore compared gene expression among HH stage 3, HH stage 10, HH stage 15, and HH stage 22. Every stage was compared to the consecutive earlier and later stage respectively. Only differentially expressed genes with a FDR < 0.01, and fold change bigger than 2 times are included for biological interpretation. (Differentially expressed gene lists are available in Additional data file 5).

1) HH stage 3 to HH stage 10

A comparison between the HH stage 3 embryo and the HH stage 10 embryo shows that only 91 genes were down-regulated, and 143 genes were up-regulated from HH stage 3 through HH stage 10. A GO term enrichment analysis of up-regulated genes in HH stage 3 show terms like “cell migration involved in gastrulation” and “blastocyst development” indicating the biological status of HH stage 3 (gastrulation). For down-regulated genes in this comparison, enriched GO terms like “skeletal system development”, “somite specification”, “heart morphogenesis”, “positive regulation of neurogenesis”, and “kidney development” (GO terms are listed in Table 1) are found.

2) HH stage 10 to HH stage 15

In total, 15 genes were significantly up-regulated in HH stage 10 compared to HH stage 15, whereas 21 genes were down-regulated in this comparison. Enriched GO terms are listed in Table 2. Enriched GO terms of up-regulated genes are “segment specification”, “androgen metabolic process”, down-regulated genes have enriched GO terms “collagen fibril organization”, “skin morphogenesis”, and “transforming growth factor beta receptor signaling pathway”.

3) HH stage 15 to HH stage 22

In this comparison between HH stage 15 and HH stage 22 embryos, 15 genes were up-regulated and 26 genes were down-regulated. Enriched GO terms are listed in Table 3.

Only terms “thyroid hormone generation” and “water transport” were enriched in HH 15 vs. HH 22 up-regulated genes. HH 15 vs. HH 22 down-regulated genes were enriched with terms like “mitotic metaphase”, “kinetochore assembly”, and “establishment of chromosome localization”.

Table 1. GO enrichment results of HH stage 3 vs. HH stage 10 (Pvalue < 0.01)

Embryo HH 3 vs. HH 10 up-regulated genes

GOBPID	Pvalue	Count	Size	Term
GO:0042074	0.000323	2	3	cell migration involved in gastrulation
GO:0019915	0.002911	2	8	lipid storage
GO:0033036	0.003066	2	8	macromolecule localization
GO:0001824	0.003718	2	9	blastocyst development

Embryo HH 3 vs. HH 10 down-regulated genes

GOBPID	Pvalue	Count	Size	Term
GO:0007275	4.92E-07	33	826	multicellular organismal development
GO:0001501	1.61E-05	8	64	skeletal system development
GO:0030199	1.93E-05	4	10	collagen fibril organization
GO:0048704	0.000117	4	15	embryonic skeletal system morphogenesis
GO:0051146	0.000198	4	17	striated muscle cell differentiation
GO:0007389	0.000228	7	70	pattern specification process
GO:0050878	0.000277	6	51	regulation of body fluid levels
GO:0021514	0.000322	2	2	ventral spinal cord interneuron differentiation
GO:0021522	0.000322	2	2	spinal cord motor neuron differentiation
GO:0048665	0.000322	2	2	neuron fate specification
GO:0006950	0.000466	8	101	response to stress
GO:0001822	0.000803	4	24	kidney development
GO:0055010	0.000847	3	11	ventricular cardiac muscle morphogenesis
GO:0001757	0.000956	2	3	somite specification
GO:0003007	0.0011	4	26	heart morphogenesis
GO:0055001	0.001114	3	12	muscle cell development
GO:0060415	0.001114	3	12	muscle tissue morphogenesis
GO:0009952	0.001948	4	31	anterior/posterior pattern formation
GO:0030168	0.002216	3	15	platelet activation
GO:0001657	0.002993	2	5	ureteric bud development
GO:0001658	0.003112	2	5	ureteric bud branching
GO:0032781	0.003112	2	5	positive regulation of ATPase activity
GO:0051592	0.003821	3	18	response to calcium ion
GO:0007517	0.003875	6	84	muscle development
GO:0043062	0.004174	4	37	extracellular structure organization
GO:0008277	0.004478	3	19	regulation of G-protein coupled receptor protein signaling pathway
GO:0006559	0.004613	2	6	L-phenylalanine catabolic process
GO:0030049	0.004613	2	6	muscle filament sliding
GO:0070252	0.004613	2	6	actin-mediated cell contraction
GO:0051960	0.006605	4	42	regulation of nervous system development
GO:0000122	0.007805	5	69	negative regulation of transcription from RNA polymerase II promoter
GO:0050769	0.008266	2	8	positive regulation of neurogenesis

Table 2. GO enrichment results of HH stage 10 vs. HH stage 15 (Pvalue < 0.01)

Embryo HH 10 vs. HH 15 up-regulated genes

GOBPID	Pvalue	Count	Size	Term
GO:0008209	0.00368	1	2	androgen metabolic process
GO:0030573	0.00368	1	2	bile acid catabolic process
GO:0006590	0.007348	1	4	thyroid hormone generation
GO:0006699	0.007348	1	4	bile acid biosynthetic process
GO:0006707	0.007348	1	4	cholesterol catabolic process
GO:0007379	0.009178	1	5	segment specification
GO:0008207	0.009178	1	5	C21-steroid hormone metabolic process

Embryo HH 10 vs. HH 15 down-regulated genes

GOBPID	Pvalue	Count	Size	Term
GO:0030199	1.27E-06	3	10	collagen fibril organization
GO:0043062	7.95E-05	3	37	extracellular structure organization
GO:0030644	0.002394	1	1	cellular chloride ion homeostasis
GO:0043206	0.002394	1	1	fibril organization
GO:0055083	0.002394	1	1	monovalent inorganic anion homeostasis
GO:0043589	0.004783	1	2	skin morphogenesis
GO:0032501	0.005823	3	303	multicellular organismal process
GO:0007179	0.006315	2	51	transforming growth factor beta receptor signaling pathway
GO:0006600	0.007166	1	3	creatine metabolic process
GO:0009650	0.007166	1	3	UV protection
GO:0030002	0.007166	1	3	cellular anion homeostasis
GO:0050777	0.007166	1	3	negative regulation of immune response
GO:0006833	0.009545	1	4	water transport

Table 3. GO enrichment results of HH stage 15 vs. HH stage 22 (Pvalue < 0.01)

Embryo HH 15 vs. HH 22 up-regulated genes				
GOBPID	Pvalue	Count	Size	Term
GO:0006590	0.007348	1	4	thyroid hormone generation
GO:0006833	0.007348	1	4	water transport

Embryo HH 15 vs. HH 22 down-regulated genes				
GOBPID	Pvalue	Count	Size	Term
GO:0000089	0.002578	1	1	mitotic metaphase
GO:0007080	0.002578	1	1	mitotic metaphase plate congression
GO:0015670	0.002578	1	1	carbon dioxide transport
GO:0032314	0.002578	1	1	regulation of Rac GTPase activity
GO:0035021	0.002578	1	1	negative regulation of Rac protein signal transduction
GO:0007079	0.00515	1	2	mitotic chromosome movement towards spindle pole
GO:0008209	0.00515	1	2	androgen metabolic process
GO:0018076	0.00515	1	2	N-terminal peptidyl-lysine acetylation
GO:0018205	0.00515	1	2	peptidyl-lysine modification
GO:0030573	0.00515	1	2	bile acid catabolic process
GO:0051058	0.00515	1	2	negative regulation of small GTPase mediated signal transduction
GO:0051382	0.00515	1	2	kinetochore assembly
GO:0051303	0.007716	1	3	establishment of chromosome localization

Comparisons of genes identified in adult stage and embryonic stages in chicken

In total, 2,476 housekeeping genes were identified being expressed in all 8 adult tissues and defined as “housekeeping genes” in Chapter 4, about 81% (2,011 out of 2,476 genes) of those housekeeping genes in adult tissues were also identified being expressed in all 4 whole embryo stages and 8 embryonic tissues. Not surprisingly, the enriched GO terms like “RNA processing”, “translation”, and “protein folding” were present in both housekeeping gene lists. In contrast, 672 genes were identified being expressed specifically in only one individual adult tissue in Chapter 4, about 13% (88 out of 672 genes) were also being expressed specifically in the same corresponding tissue types in embryonic tissues. Enriched GO terms of tissue-specific genes were similar in some tissues and are different in other tissues at different time (adult stage vs. embryonic stage). For example, the enriched GO terms (p-values < 0.01) for adult brain-specific genes and embryonic brain-specific genes were quite similar (shown in Table 4). Enriched terms related to central nervous systems like “synaptic transmission”, “learning”, and “neuron development” were present in both adult and embryonic brains. However, the enriched GO terms for intestine-specific genes were quite different comparing adult and embryonic stages (Table 5). In adult intestines, many metabolic processes were observed, including “digestion”, “proteolysis” and other terms, whereas in embryonic intestine, only very few metabolic processes were observed like “bile acid metabolic process”.

Table 4. Enriched GO terms in brain-specific genes (adult vs. embryonic stages)

Embryonic brain specific genes

GOBPID	Pvalue	Count	Size	Term
GO:0007268	6.90E-09	12	87	synaptic transmission
GO:0051179	1.70E-04	32	1098	localization
GO:0010243	2.00E-04	3	8	response to organic nitrogen
GO:0006812	2.30E-03	11	258	cation transport
GO:0001975	2.40E-03	2	5	response to amphetamine
GO:0007185	2.40E-03	2	5	transmembrane receptor protein tyrosine phosphatase signaling pathway
GO:0032990	2.80E-03	5	62	cell part morphogenesis
GO:0006835	3.50E-03	2	6	dicarboxylic acid transport
GO:0008542	3.50E-03	2	6	visual learning
GO:0000904	4.70E-03	5	70	cell morphogenesis involved in differentiation
GO:0007214	4.90E-03	2	7	gamma-aminobutyric acid signaling pathway
GO:0015813	8.30E-03	2	9	L-glutamate transport
GO:0030030	8.50E-03	6	113	cell projection organization
GO:0048666	9.70E-03	5	84	neuron development

Adult brain specific genes

GOBPID	Pvalue	Count	Size	Term
GO:0048856	6.00E-07	17	166	anatomical structure development
GO:0051179	2.00E-06	42	803	localization
GO:0022008	5.30E-06	14	136	neurogenesis
GO:0007399	1.60E-04	10	111	nervous system development
GO:0003008	2.80E-04	22	395	system process
GO:0022010	6.20E-04	2	2	myelination in the central nervous system
GO:0048667	8.20E-04	7	62	cell morphogenesis involved in neuron differentiation
GO:0030182	1.30E-03	9	109	neuron differentiation
GO:0048709	2.20E-03	3	11	oligodendrocyte differentiation
GO:0016043	2.50E-03	8	86	cellular component organization
GO:0048858	4.30E-03	6	62	cell projection morphogenesis
GO:0051649	4.50E-03	7	82	establishment of localization in cell
GO:0007275	4.60E-03	36	942	multicellular organismal development
GO:0031175	5.30E-03	6	65	neurite development
GO:0007269	6.80E-03	3	16	neurotransmitter secretion
GO:0019228	8.10E-03	3	17	regulation of action potential in neuron
GO:0007612	9.50E-03	3	18	learning

Table 5. Enriched GO terms in intestine-specific genes (adult vs. embryonic stages).

Embryonic intestine specific genes				
GOBPID	Pvalue	Count	Size	Term
GO:0008206	0.0038	2	9	bile acid metabolic process
GO:0007040	0.0057	2	11	lysosome organization
GO:0031175	0.0079	2	13	neurite development

Adult intestine specific genes				
GOBPID	Pvalue	Count	Size	Term
GO:0007586	3.40E-06	5	29	digestion
GO:0006508	1.30E-03	7	191	proteolysis
GO:0006071	5.30E-03	2	13	glycerol metabolic process
GO:0003051	8.60E-03	1	1	angiotensin-mediated drinking behavior
GO:0006005	8.60E-03	1	1	L-fucose biosynthetic process
GO:0009226	8.60E-03	1	1	nucleotide-sugar biosynthetic process
GO:0019372	8.60E-03	1	1	lipoxygenase pathway
GO:0019673	8.60E-03	1	1	GDP-mannose metabolic process
GO:0042351	8.60E-03	1	1	'de novo' GDP-L-fucose biosynthetic process
GO:0046368	8.60E-03	1	1	GDP-L-fucose metabolic process
GO:0046813	8.60E-03	1	1	virion attachment, binding of host cell surface receptor

Discussion

Gene expression distribution in embryo stages and embryonic tissues

We have defined genes being expressed in each individual stage/tissue by comparing to negative control spots on the array (Figure 1a) and observed that to a certain extent most genes show some tissue-specific expression pattern (Figure 1b). For each individual embryonic stage/tissue, specifically expressed genes were identified and these genes were used as candidate genes to study the biological processes during each of these specific stages and tissues. Our GO enrichment analyses show that many stage/tissue-specific genes were enriched with GO terms corresponding to the biological functions of tissues from which they originated. This information can be used to infer further stage/tissue-specific genes among the genes that are represented on the array by probes currently lacking any annotation. Furthermore, given the expression distribution of genes across the 12 conditions surveyed, a list of widely expressed genes (housekeeping genes) was identified. The enriched GO terms of housekeeping genes imply that cell proliferation is the universal process in all tissues during developmental stages. For example, the embryonic bursa tissue had enriched GO terms like “immunoglobulin production” indicating that bursa, as a major immune organ in adult chicken, is already functioning in early embryonic stages (HH stage 36).

Compactness of housekeeping genes during embryonic development

The genes which are expressed under all 12 conditions were defined as housekeeping genes. Similar as was observed for the housekeeping genes described in chapter 4 the gene length for the housekeeping genes identified in this study also were shorter then those of the stage/tissue specific genes. Also similar are the observed shorter coding sequence length, average intron length, and length of the intergenic region. This finding suggests a selection for compactness of widely expressed genes with universal biological functions to reduce the costs of transcription. Similar findings were reported in humans [11] and in chicken (Chapter 4). As discussed in Chapter 4, one possible scenario to explain the larger none-coding regions of tissue-specific genes would be that larger “regulatory spaces” would allow more complex regulation on transcription of those genes through a larger number of regulatory sequences.

Biological processes during embryonic development

As indicated by the enriched GO terms of housekeeping genes identified in all stages/tissues, the most clear essential biological processes during the different developmental stages are related to cell division. The GO enrichment analyses for the differentially expressed stage–specific genes identified from the pair-wise comparisons among different embryonic stages show a gradual change of development from the HH stage 3 embryos to the HH stage 22 embryos. For instance, heart morphogenesis”, “positive regulation of neurogenesis”, and “kidney development” were enriched in HH 3 vs. HH 10 down-regulated genes, these terms suggest that, based on transcriptional profiles and compared to the HH 3 stage, the development of several major organs like heart, kidney, and CNS already started to develop during these early embryonic stages. It is known that the first formation of the “head process” becomes apparent at HH stage 5 [5], and the embryonic head develops even further at HH stage 10 compared to HH stage 5. The tubular heart has completely fused at the level of the presumptive ventricle and begins to beat around HH stage 10-11 in chicken embryos [12]. The enriched GO terms identified in differentially expressed genes between HH stage 10 comparing to HH stage 3 embryos provide crude pictures of embryonic development, sometimes on surprisingly detailed levels about individual organ development. The presumptive skins in HH stage 10 embryos have been reported previously, and the development of embryonic skin continues during embryonic development to later stages [13]. The enriched terms “collagen fibril organization” and “skin morphogenesis” were found in HH10 vs. HH15 down-regulated genes and this suggests that the changes of embryonic skin development from HH stage 10 though HH stage 15 embryos are relatively large. The GO terms related to cell division enriched in HH 15 vs. HH 22 down-regulated genes imply that the size of embryo keeps on elongating and expanding from stage HH15 to HH22, which is in agreement with earlier findings in chicken embryonic development [5].

The analyses described above show that transcriptomic approaches can detect, sometimes subtle, changes of biological processes from one stage to another. This approach is very powerful and can result in a better understanding of embryonic development in chicken.

Comparisons of genes identified in adult stage and embryonic stages in chicken

The largest proportion of the housekeeping genes identified in adult tissues was also identified as housekeeping genes in embryonic stages/tissues. This implies the housekeeping genes identified in Chapter 4 and Chapter 5 were indeed involved in essential biological processes to maintain normal functions of living cells in different tissues at different times during development. Furthermore, only a very small proportion of the tissue-specific genes identified in adult tissues was also identified to be tissue-specific and to be expressed in embryonic tissues. This implies that the overall biological processes in adult tissues in general are very different compared to the corresponding embryonic tissues. The major exception to this observation is the brain. For brain-specific genes, the enriched GO terms were very much related to neurological functions in both embryonic and adult brains, indicating that the embryonic brain at HH stage 36 already has many of the basic functions of the adult brain. In contrast, for intestine-specific genes, the enriched GO terms were not very similar. Many more digestive and metabolic processes related terms were over-presented in adult intestines as compared to embryonic intestines. This most likely reflects the fact that most nutritional supplies for chicken embryos were derived from the egg yolk and that the embryonic intestine is not yet completely functioning as a digestive organs., Studying tissues at different developmental stages using a transcriptomic approach can provide a global picture of the biological processes in different organs and provide further knowledge within regards the biological functions of the different organs at different developmental stages.

Conclusions

We have used microarrays to survey gene expression across 4 stages of whole embryos and 8 different embryonic tissues and identified stage/tissue-specific genes and housekeeping genes and studied their functions by testing enriched GO terms. Compactness of housekeeping genes was shown indicating a selection on economy for transcription in widely expressed genes. Furthermore, expression levels between different stages of whole embryos were compared and the enriched GO terms reflected the changes in biological processes from one stage of embryonic development to the next. Our dataset provides a unique resource for further studies on molecular mechanism during chicken embryo development and embryonic organ functions in the future.

Materials and Methods

Embryo/tissue sample preparation

All the embryo/embryonic tissue samples were collected and immediately put into RNAlater RNA Stabilization Reagent (Qiagen, Valencia, CA, USA), followed by incubation at +4℃ overnight, then stored at -80℃ until use. Sev eral individual embryos or embryonic tissues were pooled together to get enough quantity of RNA for individual hybridization. Detailed information about the samples is shown in Table 6.

Table 6. An overview of embryonic stages/tissue samples included in this study.

HH stage	tissue	description
HH 3+	whole embryo	HH stage 3+, 12 hours after incubation
HH 10	whole embryo	HH stage 10, 30-48 hours after incubation
HH 15	whole embryo	HH stage 15+, 55 hours after incubation
HH 22	whole embryo	HH stage 22, >72 hours after incubation
HH 36	Brain	after 10 days incubation
	Bursa	after 10 days incubation
	Heart	after 10 days incubation
	Kidney	after 10 days incubation
	Liver	after 10 days incubation
	Lung	after 10 days incubation
	Intestine	after 10 days incubation
	Spleen	after 10 days incubation

RNA isolation, labeling and hybridization

Total RNA was isolated using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer’s instructions, followed by a subsequent sample "clean-up" using RNeasy Mini Kit (Qiagen, Valencia, CA, USA). RNA quantity was measured using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, USA). The quality and integrity of the RNA was analyzed using the Agilent Bioanalyzer 2100 (Palo Alto, CA, USA), RNA was amplified using MessageAmp™ II aRNA Kit (Ambion, Foster City, CA, USA) and cRNA was further used for chemical coupling with ULS-Cy3/Cy5 (ULS™ aRNA labeling kit; Kreatech, Amsterdam, Netherlands). After coupling and purification the cRNA concentration and fluorescent incorporation was quantified using the Nanodrop Spectrophotometer. One µg of each labeled cRNA was used to hybridize on the Ark-Genomics G.gallus 20k array. The hybridizations were done overnight on a GeneTAC hybridization station (Genomic Solutions, Holliston, MA, USA). Hybridized arrays were scanned using Agilent DNA microarray scanner (Agilent, Santa Clara CA, USA). A common reference design was used in this study. The common reference was made by pooling total RNA samples from all individual samples, and each individual sample was hybridized against the common reference on the same array slide. In all 40 arrays (5 biological replicates per tissue type), cRNA of individual tissue samples was always labeled using Cy3 (green), and the cRNA of the common reference samples were always labeled using Cy5 (red).

Array probe re-annotation

The ARK-Genomics G. gallus 20K array platform, used in this study, contains 20,460 unique oligonucleotide probes (NCBI GEO [14] accession GPL5480). The probe sequences were mapped to the current chicken genome assembly (Ensembl Genome database release 50, WASHU2 assembly, May 2006) using the method described previously by Neerincx et al. [6]. In total, 14,900 probes were mapped to a unique position in the current assembly, corresponding to 8,792 unique ensembl genes in the Ensembl Genome Database and 5,357 probes were mapped to unique positions in the genome without a link to an ensembl gene.

Microarray data processing, normalization, and statistical analysis

Scanned TIFF images were analyzed using GenePix 6.0 (Axon, Sunnyvale, CA, USA), and results were saved as GenePix Result (*.gpr) files. We used R/Bioconductor package Limma [15] to analyze the array data. The *.gpr files were imported into R [16] (version 2.8.0), median values of both foreground and background intensities were extracted and used in the following analysis. We gave any spot with FLAG-value less than -50 (these spots were flagged as “bad spot” by GenePix program or manually) a weight of 0.01, and all the other spots we gave weights of 1. We used background correction option “normexp+offset” (offset=50) [17], background corrected data were normalized using “printtiploess” normalization (within array normalization) followed by “Rquantile” normalization implemented in the Limma package.

For the purpose of defining a gene being expressed, the normalized intensities of the green channel (representing all individual tissue samples) were used as gene expression data in the analysis and the data points for those spots (both genes and negative controls) with low weight (0.01) were removed in further analysis. The gene expression data was first averaged within each tissue type among the five biological replicates, and then the gene expression data for probes targeting the same Ensembl genes/entrez gene were averaged.

For the purpose of identifying differentially expressed gene, normalized log ratio (log2(R/G)) data were used, all data for control spots on the array were removed before fitting in the linear model, probe data was used for differential expression analysis and differential expression of individual genes was assessed using linear modeling and empirical Bayes methods [18] as implemented in the R package Limma. Multiple testing was corrected using the False Discovery Rate (FDR) control method described by Benjamini and Hochberg [19]. Only probes with a FDR < 0.01 and a fold change bigger than 2 are included for biological interpretation in this study.

Gene Ontology term enrichment analysis

All the genes having a chicken Ensembl gene ID were mapped to their 1-to-1 human orthologous genes using Bioconductor package biomaRt [20] through the Ensembl Genome Database. The GO term enrichment analysis was subsequently performed using human gene annotation using R package GOstats. A conditional hypergeometric test algorithm provided within GOstats package was applied to GO enrichment analysis. The conditional hypergeometric test identifies a GO term as significant if there is evidence beyond that provided by its significant children. Only the enriched GOBP terms with raw p-values < 0.01 were used for biological interpretation in this study.

Abbreviations

FDR: False Discovery Rate;
GO: Gene Ontology;
CNS: Central Nervous System;

Authors' contributions

HN designed and analyzed the microarray study, and drafted the manuscript. HN, IMLJ, MAGB, MKR, RPMAC were involved in embryo staging and sampling, MAMG supervised the study and all authors helped to revise the manuscript. All authors have red and proved the final manuscript.

Acknowledgements

This work was financed by European Union grant FOOD-CT-2004—506416 (EADGENE).

References

1. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.
2. Li X, Chiang HI, Zhu J, Dowd SE, Zhou H: **Characterization of a newly developed chicken 44K Agilent microarray**. *BMC Genomics* 2008, **9**:60.
3. **ARK-Genomics Chicken 20K Oligo Array**: [http://www.arkgenomics.org/microarrays/]
4. Hamburger V, Hamilton HL: **A series of normal stages in the development of the chick embryo**. 1951. *Dev Dyn* 1992, **195**(4):231-272.
5. Mason I: **The avian embryo: an overview**. *Methods in molecular biology* (Clifton, NJ) 2008, **461**:223-230.
6. Neerincx PB, Rauwerda H, Nie H, Groenen MA, Breit TM, Leunissen JA: **OligoRAP - an Oligo Re-Annotation Pipeline to improve annotation and estimate target specificity**. *BMC Proc* 2009, **3 Suppl 4**:S4.
7. **Ensembl Genome Database**: [http://www.ensembl.org/]
8. **Entrez Gene**: [http://www.ncbi.nlm.nih.gov/gene]
9. **Gene Ontology website**: [www.geneontology.org]
10. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association**. *Bioinformatics* 2007, **23**(2):257-258.
11. Eisenberg E, Levanon EY: **Human housekeeping genes are compact**. *Trends Genet* 2003, **19**(7):362-365.
12. Martinsen BJ: **Reference guide to the stages of chick heart embryology**. *Dev Dyn* 2005, **233**(4):1217-1237.
13. Chuong CM, Jung HS, Noden D, Widelitz RB: **Lineage and pluripotentiality of epithelial precursor cells in developing chicken skin**. *Biochemistry and cell biology*

1998, **76**(6):1069-1077.

14. **National Center for Biotechnology Information (NCBI) Gene Expression Omnibus**: [http://www.ncbi.nlm.nih.gov/geo/]
15. Smyth GK: **Limma: linear models for microarray data**. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York: Springer; 2005:397-420.
16. R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. .
17. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK: **A comparison of background correction methods for two-colour microarrays**. *Bioinformatics* 2007, **23**(20):2700-2707.
18. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol*2004, **3**:Article 3.
19. Benjamini Y and Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society*, 1995, **Series B 57**:289–300.
20. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis**. *Bioinformatics* 2005, **21**(16):3439-3440.

Additional data files

Additional data file 1: Expression data of 8908 genes across 12 embryonic stages/tissues
http://abgc.asg.wur.nl/nie/Chapter_5/Additional%20data%20file%201.xls

Additional data file 2: Lists of stage/tissue-specific genes and housekeeping genes
http://abgc.asg.wur.nl/nie/Chapter_5/Additional%20data%20file%202.xls

Additional data file 3: GO enrichment analysis results for embryonic stage/tissue-specific genes (P value < 0.01)
http://abgc.asg.wur.nl/nie/Chapter_5/Additional%20data%20file%203.xls

Additional data file 4: GO enrichment analysis results for housekeeping genes (P value < 0.01)
http://abgc.asg.wur.nl/nie/Chapter_5/Additional%20data%20file%204.xls

Additional data file 5: List of differentially expression genes in pair-wise comparisons among different developmental stages (HH 3, HH 10, HH 15, HH22)
http://abgc.asg.wur.nl/nie/Chapter_5/Additional%20data%20file%205.xls

Chapter 6

Regional regulation of transcription in the chicken genome

Haisheng Nie, Richard PMA Crooijmans, John WM Bastiaansen,
Hendrik-Jan Megens, and Martien AM Groenen

Animal Breeding and Genomics Centre
Wageningen University
The Netherlands

Published in *BMC Genomic* 2010, 11: 28 (Featured article)

Abstract

Background

Over the past years, the relationship between gene transcription and chromosomal location has been studied in a number of different vertebrate genomes. Regional differences in gene expression have been found in several different species. The chicken genome, as the closest sequenced genome relative to mammals, is an important resource for investigating regional effects on transcription in birds and studying the regional dynamics of chromosome evolution by comparative analysis.

Results

We used gene expression data to survey eight chicken tissues and create transcriptome maps for all chicken chromosomes. The results reveal the presence of two distinct types of chromosomal regions characterized by clusters of highly or lowly expressed genes. Furthermore, these regions correlate highly with a number of genome characteristics. Regions with clusters of highly expressed genes have higher gene densities, shorter genes, shorter average intron and higher GC content compared to regions with clusters of lowly expressed genes. A comparative analysis between the chicken and human transcriptome maps constructed using similar panels of tissues suggests that the regions with clusters of highly expressed genes are relatively conserved between the two genomes.

Conclusions

Our results revealed the presence of a higher order organization of the chicken genome that affects gene expression, confirming similar observations in other species. These results will aid in the further understanding of the regional dynamics of chromosome evolution.

The microarray data used in this analysis have been submitted to NCBI GEO database under accession number GSE17108. The reviewer access link is: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=tjwjpscyceqawjk&acc=GSE17108

Background

Gene expression in eukaryotes is regulated on two different levels, i.e. individual gene level and regional level in the genome. The best studied, and generally considered the major level of regulation, is the regulation at the level of individual genes. Although a number of well studied exceptions have identified a number of tightly co-regulated gene clusters, such as the globin, MHC and the Hox gene gene clusters [1-4], the common model for eukaryotic gene transcription involves the binding of several transcription factors (TFs) to promoter regions and enhancers, resulting in activation of the individual genes. It has become increasingly evident that in addition to gene regulation by TF binding to regulatory sequences, eukaryotic gene expression is also regulated at a higher level, and several studies have demonstrated the dependency of gene expression on the location of the gene within the genome [5-7].

Over the past years, the relationship between gene transcription and chromosomal location has been studied in a number of different vertebrate genomes. Analysis of the human transcriptome map based on SAGE (serial analysis of gene expression) data from 12 human tissues [8] revealed the clustering of highly expressed genes within specific chromosomal regions; these regions were termed “RIDGEs”, or “Regions of Increased Gene Expression”. Genomic regions containing genes expressed at much lower levels were termed anti-RIDGEs, and these regions exhibit characteristics opposite those of RIDGEs [8, 9]. A similar region-wide regulation of gene expression was later reported in the Drosophila genome [10, 11]. RIDGEs were also found in the mouse genome [12] and are reported to be relatively conserved between the mouse and human genome [13]. A later study [14] showed gene expression to be regulated at a region-wide level in the human genome. Insertion of green fluorescent protein (GFP) reporter constructs at 90 different chromosomal positions in the human genome showed that gene transcription was regulated through a novel region-wide regulatory mechanism as well as via specific transcription factors, thereby demonstrating dual mechanisms in the regulation of gene transcription.

Regional differences in gene expression have been found in two distinct clades (mammals and flies) of the metazoan phylogeny, suggesting a common mechanism of regulation of transcription in all animals. Other characteristics of eukaryotic genomes such as gene density and recombination have also been implied to exhibit domain-like features [15]. In addition, levels of gene expression have been found to correlate with time of chromatin replication during the cell cycle, i.e. the early replication of actively expressed regions of the genome [15]. Striking in this respect is the observed location of gene-dense and highly expressed chromosomes towards the center of the nucleus and the location of gene-poor and weakly expressed chromosomes towards the nuclear envelope in both human [16] and chicken cells

[17]. Furthermore, in chicken, this spatial organization seems to correlate with chromosome size [17].

The chicken genome sequence, published in 2004, was the first non-mammalian amniote genome to become available [18]; its karyotype (2n = 78) consists of 38 autosomes and one pair of sex chromosomes, with the female being the heterogametic sex (ZW female, ZZ male). Thus far, there are 31 known chromosomes assembled in the chicken genome, including six macro-chromosomes (GGA1-5, Z), five intermediate-chromosomes (GGA6-10) and twenty micro-chromosomes (GGA11-28, 32, W) [18]. The existence of micro-chromosomes is one of the interesting features of the chicken genome [19], micro-chromosomes are also found in some primitive amphibians [20, 21] and most reptiles [22]. Besides the huge differences on sizes, microchromosomes also exhibit higher gene density, smaller gene size, and higher recombination rates compared with those in macrochromosomes [18, 23]. As the best-studied bird genome currently available, and the closest sequenced genome relative to mammals, the chicken genome is an important resource for comparative genomics, including comparative studies on gene transcription.

To investigate regional effects on transcription in birds, we analyzed chicken gene expression data across a number of different tissues to address three major questions: (i) if there are regional differences in the regulation of transcription in the chicken genome, (ii) if these regions are conserved during evolution, and (iii) the characteristics of these genomic regions in the chicken.

Results

Gene expression data

Eight different chicken tissues were used for the analysis of whole genome gene expression profiles using chicken 20k oligonucleotide microarrays (GEO [24] accession GPL8861, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=tjwjpscyceqawjk&acc=GPL8861). All array probes were designed from known transcripts and ESTs based on the chicken genome assembly WASHUC1 (Dec. 2004), and a stringent selection of probes was performed before the analysis. A total of 7477 probes failed to map to unique chicken Ensembl genes, and these were excluded to avoid the introduction of additional noise into the analysis. In total, 11,361 chicken Ensembl gene IDs located on 27 chromosomes were included in the expression study. These 27 chromosomes cover over 90% of the chicken genome, and include all macro-chromosomes and many of the micro-chromosomes. The number of Ensembl genes on each of these chromosomes is shown in Figure 1. On average, about

70% of all the known ensemble genes on each of these 27 chromosomes were included in this analysis.

In this study, we define the chicken transcriptome map as the median expression levels of the 11,361 chicken Ensembl genes across eight tissues on 27 chromosomes. The start position of the first Ensembl gene and the end position of the last Ensembl gene on each chromosome were considered the start and end of each chicken chromosome. The combined size of the chromosomal sequences analyzed in this study is 1,022,830,111 bp, which covers 97% of the total length of build 2 (WASHUC2, May 2006) of the chicken (*Gallus gallus*) genome.

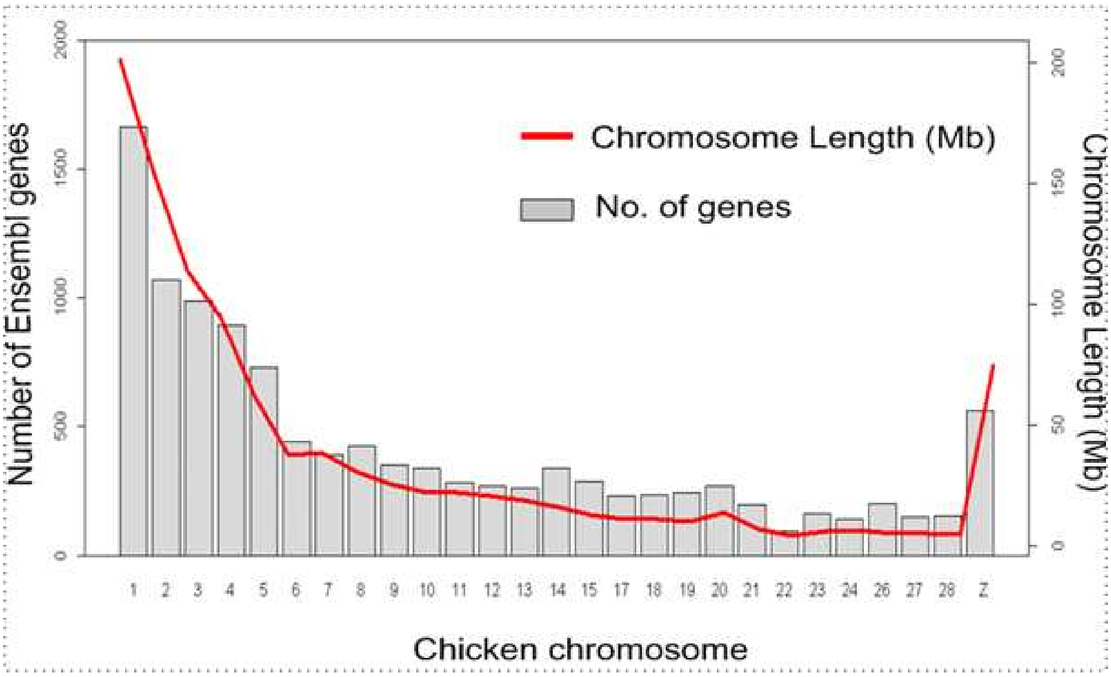


Figure 1. Distribution of genes on individual chicken chromosomes. The number of Ensembl genes on each chicken chromosome used in the analysis is shown on the y-axis on the left; the y-axis on the right shows the size of the individual chromosomes.

Regional differences of transcription in the chicken genome

To create the chicken transcriptome map, the Ensembl genes were ordered based on the middle positions of the genes on each chromosome, and a robust scatter plot smoothing (running median) technique was applied to the median expression values of the genes on each chromosome (see Materials and Methods for details). The resulting transcriptome map revealed clusters of highly expressed genes on all chicken chromosomes (Figure 2). Marked differences were observed in the overall expression levels of the different chicken chromosomes, with GGA 2, GGA14 and GGAZ showing relatively lower overall gene expression compared to the other chromosomes. Furthermore, the gene expression levels of the micro-chromosomes were observed to be higher than those of intermediate- and macro-chromosomes; the median expression level of each chromosome was observed to decrease with increased chromosome size (Figure 3). Interestingly, the sex chromosome GGAZ shows an extremely low median expression level.

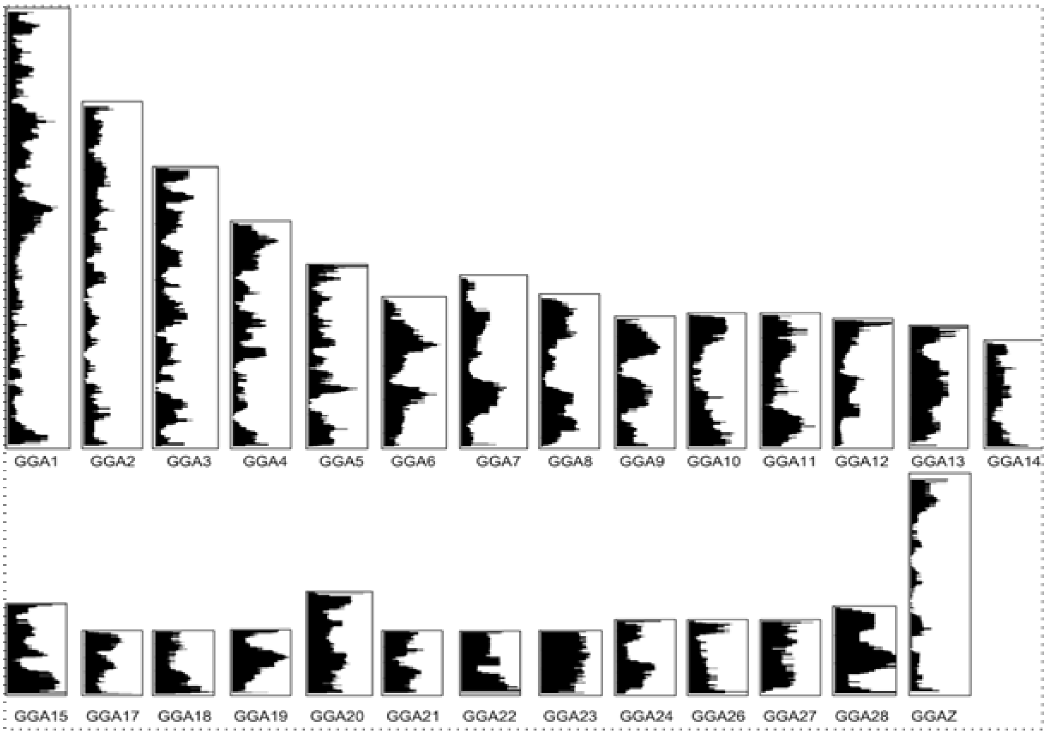


Figure 2. Regional clusters of highly expressed genes in the chicken genome. Gene expression is plotted for chicken chromosomes 1-15, 17-24, 26-28, and Z. The expression values are plotted as a moving window with a size of 39 genes to calculate the running median along the chromosomes. The log2 transformed intensities of green channel are shown; the start of the chromosomes corresponds with the top of the plot, and the window width indicates the expression levels, ranging between 6.6-8.3 (log2 scale).

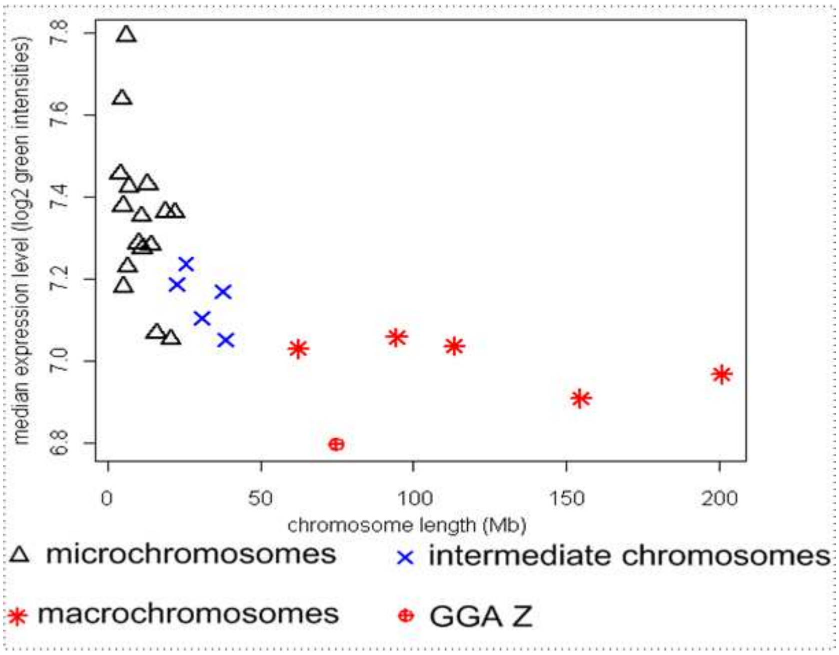


Figure 3. Relationship between median expression levels and chromosome length (correlation = -0.67, Pearson correlation).

To further investigate the unequal distribution of gene transcription activity along chicken chromosomes, we selected regions with clusters of the most highly expressed genes and regions with clusters of most lowly expressed genes, such that each region type covered approximately ten percent of the chicken genome. To be consistent with previous studies in humans [8, 9], here we use the terms "RIDGE" and "anti-RIDGE" to refer to regions showing the highest and lowest expression levels, respectively, in the chicken genome. Similar to Caron et al. [8], we define RIDGEs in the chicken genome as genomic regions with at least 10 consecutive running medians larger than 1.19 times the median expression of the chicken transcriptome, i.e. all 11,361 Ensembl genes. With a running median of a window size of 39 genes, we identified 64 RIDGEs in the chicken genome that cover approximately 10% of the genome. Using the same window size, we identified 27 anti-RIDGEs, which cover approximately 10% of the chicken genome; these anti-RIDGEs are defined as genomic regions with at least 10 consecutive running medians smaller than 0.78 times the median expression of the chicken transcriptome. The total number of Ensembl genes located in RIDGEs and anti-RIDGEs is 3260 and 1051, respectively. The mean of the median expression values of genes located in RIDGEs across the tissue panel is approximately 1.8 times higher than that of genes in anti-RIDGEs (Additional data file 1). More detailed information of RIDGEs and anti-RIDGEs can be found in Additional data file 1.

The distribution of the expression of the genes located in RIDGEs and anti-RIDGEs is shown in Figure 4. The majority of genes in anti-RIDGEs is below 7 (the log2 transformed intensities of the green channel). This is in strong contrast with the distribution observed for RIDGEs, which show a much broader distribution; furthermore, the majority of genes in RIDGEs show an expression above 7 (the log2 transformed intensities of the green channel).

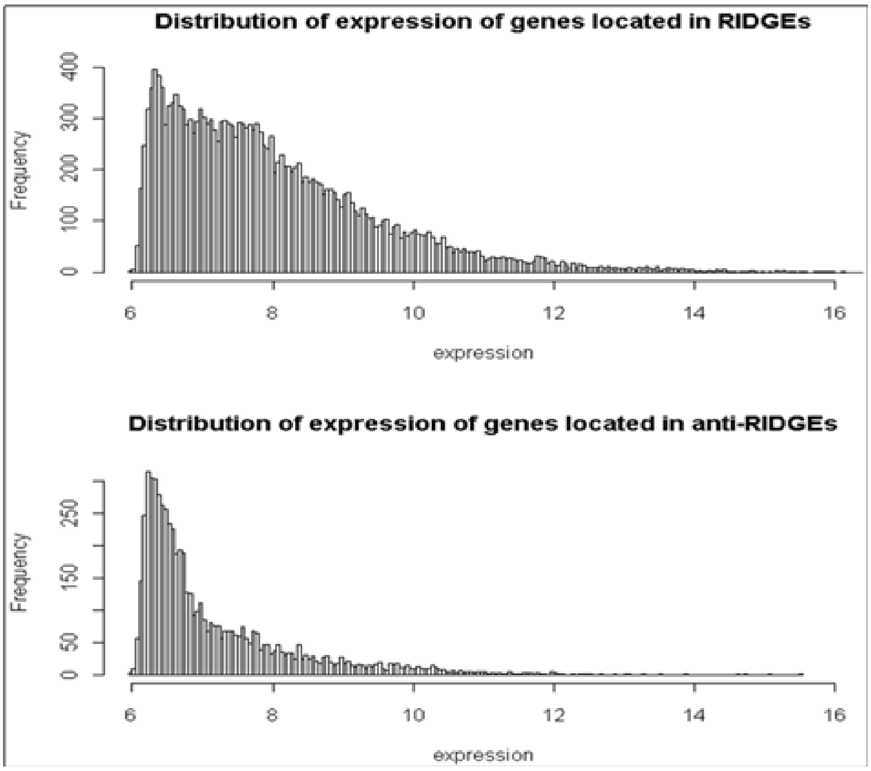


Figure 4. Histograms of gene expression values across 8 tissues for genes in RIDGEs and anti-RIDGES. Gene expression on the x-axis is the log2 transformed intensity of the green channel.

Transcriptome maps in different tissues are highly correlated

To next evaluate transcriptome maps of different types of tissues, we created transcriptome maps for each individual tissue type by applying a running median on expression values within each tissue using a window size of 39 genes. Chromosome 1 is shown in Figure 5 as an example, and the transcriptome maps for the different tissues were observed to be very similar. We performed a correlation test between the transcriptome map created using the median expression values across the eight tissues and the transcriptome maps created using the expression values from each tissue type. All transcriptome maps are highly correlated, with an average correlation of 0.88. All pair-wise correlations were highly significant, with p-values less than 2.2 x 10-16. (All pair-wise correlations between the tissue-specific transcriptome maps are shown in Additional data file 2).

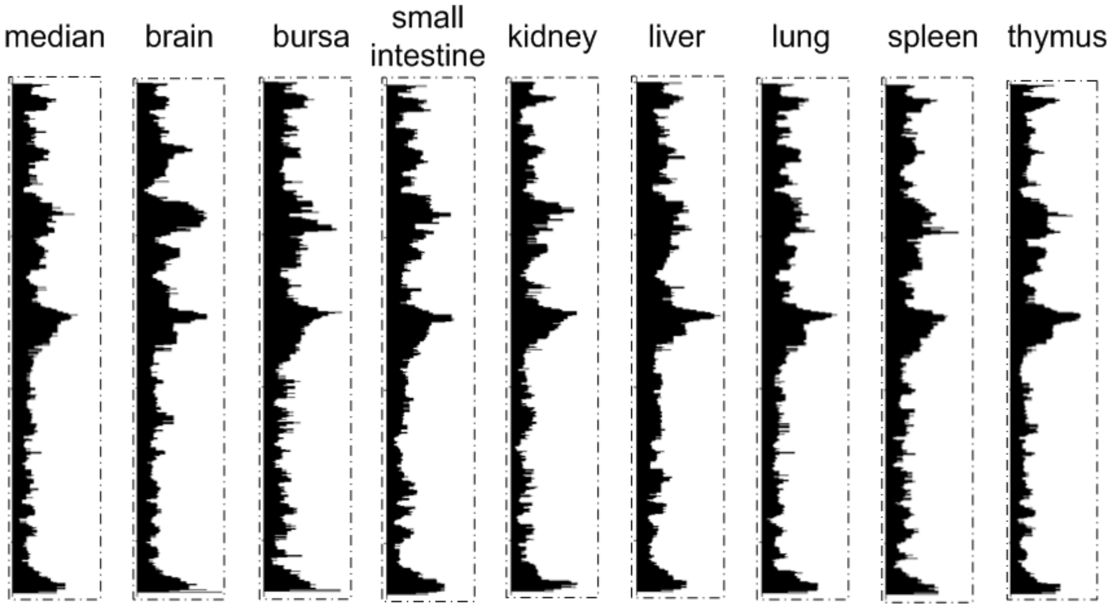


Figure 5. Transcriptome maps of chromosome 1 for different tissue types, the expression values are plotted as a moving window with a size of 39 genes to calculate the running median along the chicken chromosome 1. the start of the chromosomes corresponds with the top of the plot, and the window width indicates the expression levels, ranging between 6.6-8.3 (log2 scale).

Random permutation tests of RIDGE identification

To test the significance of the number of RIDGEs identified in our analysis, we performed random permutation tests using the same window size and threshold for RIDGE identification. In total, 10,000 random transcriptome maps were generated by permutating the gene orders throughout the genome. The permutation tests, shown in Additional data file 3, clearly show that the number of RIDGEs identified in our analysis is higher than would have been expected merely by chance.

RIDGEs are relatively conserved between chicken and human

The observation that highly expressed genes tend to be clustered within RIDGEs in the chicken as well as the human genome suggests a conserved functional organization of the genome of these vertebrates. We therefore decided to assess whether genes in RIDGEs remain associated during evolution. Thus, we consider two different forms of functional constraint. The first possibility is that specific genes within a particular RIDGE need to be co-regulated; in this case, one would expect relatively few syntenic breaks to occur within the RIDGEs. The other possibility is that genes do not need to co-localize with specific genes, but rather remain spatially associated with other highly expressed genes in general. In this case, one would expect syntenic breaks to occur specifically between two different RIDGEs. Random rearrangements of RIDGEs and anti-RIDGEs, on the other hand, would reduce the clustering of genes, and therefore abolish the effect of regional regulation of transcription. First we tested if the observed RIDGEs were less prone to be broken down during evolution from chicken to human. Previous studies comparing the human, mouse, rat, and chicken genomes identified a total of 586 conserved synteny blocks [25]. Because the identification of these synteny blocks was based on chicken genome assembly WASHUC1 (Dec. 2004), we mapped the ends of these syntenic blocks to the current chicken genome assembly (WASHUC2, May 2006) (Additional data file 4), and considered each end as an evolutionary break point. In total, we mapped 1130 break points on the WASHUC2 chicken genome assembly; we found 253 break points within RIDGEs, and 50 break points within anti-RIDGEs. Chi-square tests showed a significantly higher average number of break points in RIDGEs compared to regions outside RIDGEs (p value < 2.2x10-16) and a significantly lower number of break points in anti-RIDGEs compared to regions outside anti-RIDGEs (p value=4.18x10-10) (Additional data file 5).

To compare the transcriptome maps between chicken and human, we downloaded human gene expression data for the same types of tissues (see Materials and Methods) from the Human Transcriptome Map website [26]. Using the median of the expression values across the seven human tissues for each human gene, we performed an identical analysis on the human data as the chicken expression data to identify RIDGEs and anti-RIDGEs in the human genome. Similar to the chicken, in the human genome, RIDGEs and anti-RIDGEs each cover about ten percent of the genome. Defining the syntenic break points in the human genome using data described by Bourque et al. [25], we found a total of 143 and 86 break points in RIDGEs and anti-RIDGEs, respectively. Again, similar to results seen in the chicken, chi-square tests show a higher average number of break points in RIDGEs compared to regions outside of RIDGEs (p value=0.01) and a lower number of break points in anti-RIDGEs compared to outside anti-RIDGEs (p value=0.002) (Additional data file 5).

We identified 46 RIDGE-to-RIDGE break points and 11 anti-RIDGE-to-anti-RIDGE break points between the chicken and human genomes. Chi-square tests showed a significantly higher number of RIDGE-to-RIDGE break points between the chicken and human genomes (p value<2.2x10-16) compared to that expected by chance, and no significant difference in the number of anti-RIDGE-to-anti-RIDGE break points (p value=0.8).

Genomic characteristics of RIDGEs and anti-RIDGEs in chicken

Next we evaluated whether RIDGEs and anti-RIDGEs were associated with other genome characteristics. Positive correlations were found between chicken transcriptome map and gene density (p value<2.2x10-16), GC content (p value<2.2x10-16) and average intron length (p value<2.2x10-16). As an example, the whole chromosome views of the transcriptome map, gene density, GC content, gene length, average intron length and recombination rate are shown for chromosome 1 (Figure 6); these various parameters were similar in RIDGEs and anti-RIDGEs. To further investigate the specific genomic characteristics of RIDGEs and anti-RIDGEs, we compared the average intron length (averaged intron length of all transcripts per gene), gene length (genomic length), gene density (number of genes per 100 kb), and GC content between genes located in RIDGEs and anti-RIDGEs (Figure 7). Compared to the entire chicken genome, RIDGEs, on average, harbor genes with shorter average intron length (p value<2.2x10-16), shorter gene length (p value<2.2x10-16), and a higher GC content (p value<2.2x10-16). Anti-RIDGEs, on the other hand, show opposite trends, with genes with longer average intron length (p value<2.2x10-16), longer gene length (p value<2.2x10-16), and lower GC content (p value<2.2x10-16). Furthermore, RIDGEs also have a significantly higher gene density (p value=1.29x10-9) than anti-RIDGEs.

Gene Ontology term enrichment analysis for genes in RIDGEs and anti-RIDGEs

Our results indicate that RIDGEs are relatively conserved between human and chicken. Assuming RIDGEs are the result of evolutionary events favoring the clustering of genes with higher expression levels, one can hypothesize that genes within RIDGEs may share similar functions or biological pathways. To investigate this possibility, we performed Gene Ontology (GO) [27] term enrichment analysis on genes located in RIDGEs and anti-RIDGEs using R package Gostats [28]. However, no significant GO_BP terms (the minimum FDR of all three tests is 0.4) were found for genes in RIDGEs and anti-RIDGEs after correcting for multiple testing (Additional data file 6).

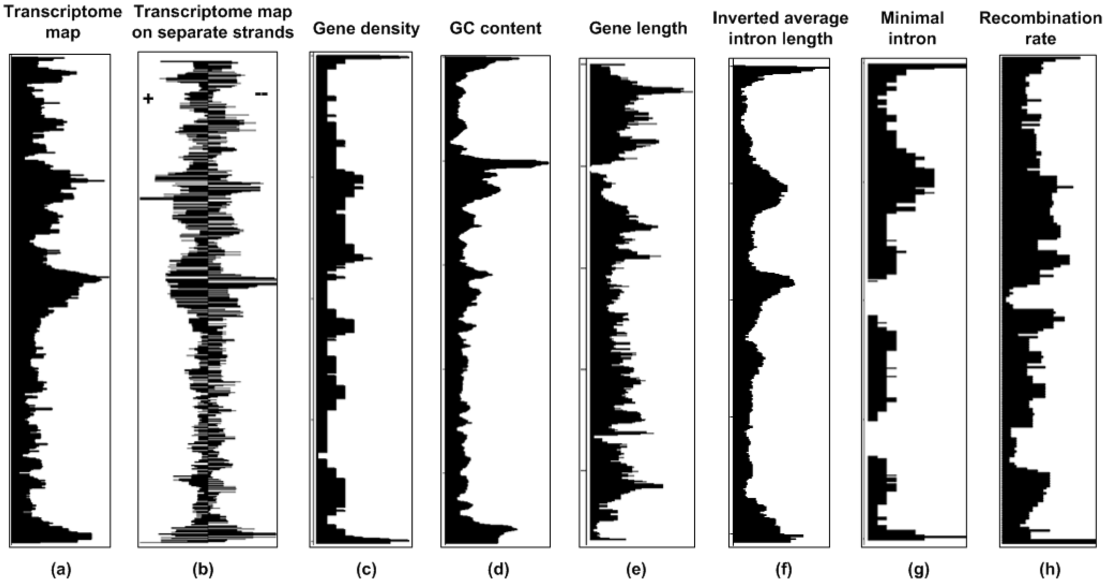


Figure 6. Whole-chromosome view of **(a) transcriptome map** (plotting running medians of gene expression values along chromosome 1 with window size of 39 genes); **(b) transcriptome map on separate strands** (plotting running medians of gene expression values on separate strands with window size of 19 genes on each individual strand (left side: + strand; right side: - strand) along chromosome 1); **(c) gene density** (gene density was defined as number of genes per 100 kb genomic region, running medians of gene densities with window size 39 gene were plotted along chromosome 1) ; **(d) GC content**, **(e) gene length**, **(f) average intron length** (GC content, gene length, and average intron length were calculated for each gene, the running medians of values for those three features with a window size of 39 genes were plotted along chromosome 1), **(g) “minimal intron” density** (the minimal intron here were defined as introns sizing from 50 to 150 bp, and minimal intron density was defined as the number of minimal introns per 500 kb genomic region, then the running medians of minimal intron intensities with window size of 39 genes were plotted along chromosome 1); and **(h) recombination rate** (recombination rate data of chicken chromosome 1 was obtained from previous study by Groenen et al.[25], and plotted in the same way as described by Groenen et al.) plotted on chicken chromosome one. The start of the chromosome corresponds with the top of the plot.

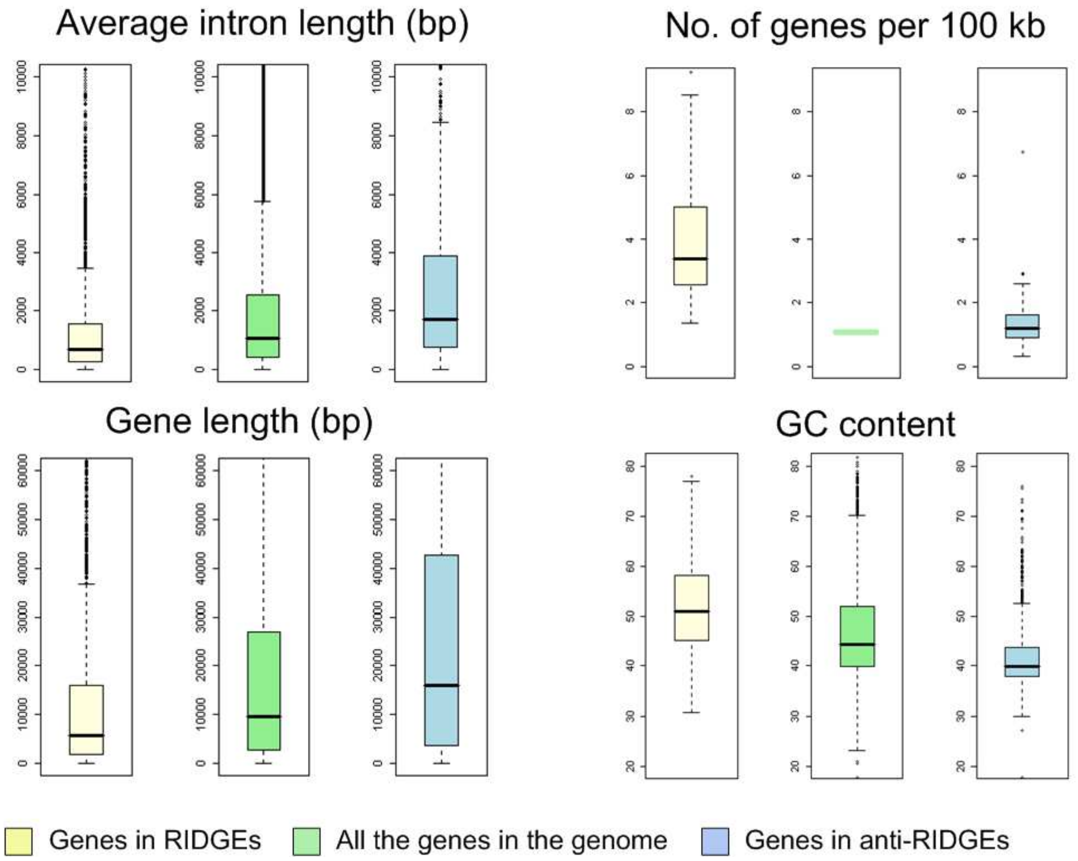


Figure 7. Boxplot of average intron length, gene length, gene density (number of genes per 100 kb) and GC content for genes in RIDGEs, anti-RIDGEs, and the complete chicken genome. The middle line of each box represents the median values. The edges of each box represent the first and third quartile values.

Discussion

Gene expression data

The annotated genes on the array platform used in this study cover most of the current chicken genome assembly. The number of genes analyzed on each chromosome is also in good proportion with chromosome length (Figure 1), which suggests against a bias in the analysis due to uneven distribution of the genes in the chicken genome. We chose to exclude chromosome 16 and 25 from our analysis, as only 24 and 59 Ensembl genes are represented on the array; this number is too low to identify any meaningful high or low expressing regions with the window size of 39 genes used in this analysis.

No major effect of different tissues on chicken transcriptome map

We observed high correlations (average correlation=0.88) among the different transcriptome maps based on the expression data from the eight different individual tissues as well as between these transcriptome maps and the transcriptome map of the combined expression data of all eight tissues. This indicates that use of the median expression value or the expression values from individual tissues only has a minor effect on the transcriptome maps and on the identification of RIDGEs and anti-RIDGEs. This shows that regional differences in transcription are a general trend in the chicken genome, even among different tissue types.

Regional differences of transcription in the genome

This is the first study in birds to construct a transcriptome map and to confirm the existence of regional differences on transcription regulation in the chicken genome. RIDGEs have been discovered in several animal species from phylogenetically distinct groups, suggesting that the existence of RIDGEs may be universal in the animal kingdom [8, 10-14].

Gierman et al. [14] showed that RIDGEs are may contain up to 80 genes and can exert an eightfold difference on the expression levels of integrated genes. They found that gene expression levels are not highly correlated to adjacent genes, but instead more correlated to the entire block of up to 80 genes, demonstrating regional effects on gene transcription. The exact mechanism underlying how gene expression occurs in RIDGEs is still unknown. One hypothesis is that evolution favors highly expressed genes to be physically close to each other, as transcription of one gene would help the chromatin of neighboring genes to “open up” during transcription. This hypothesis is in agreement with our observation of no apparent evolutionary constraint on the co-localization of specific genes, whereas we observed specific localization of specific genes within RIDGEs (see below). Goetze et al. [29] showed that RIDGEs in general are less condensed, more irregularly shaped, and are located more closely to the nuclear center than anti-RIDGEs. Furthermore, the chromatin structures of RIDGEs and anti-RIDGEs are largely independent of tissue-specific variations in gene expression and differentiation state. Their discovery again confirms the hypothesis that the different regional effect of gene transcription in RIDGEs and anti-RIDGEs is, at least in part, explained by the chromatin structure of the two types of genomic regions.

Genomic Characteristics in RIDGEs and anti-RIDGEs in chicken

Many studies have shown that chicken genome characteristics such as recombination frequency, gene density and GC density correlate with chromosome size [18, 23]. Our results show a similar trend with regard to the level of gene expression and density of RIDGEs. In the chicken, the median expression values decrease with increased chromosome length (Figure 3), which can only be partly explained by the higher gene density of the micro-chromosomes. Our permutation analysis clearly shows that the organization of genes in clusters of highly expressed genes is not random and suggests a functional mechanism. This is further strengthened by our observation that the same distribution of RIDGEs is seen when both strands of the same chromosome are analyzed separately (Figure 6). This is additional confirmation of region-like regulation of transcription during gene expression, since the opening of chromatin structures during gene expression will affect both strands by facilitating the access of transcription factors to target genes, thus enhancing gene expression in that region. Furthermore, we also found a correlation between the transcriptome maps and gene density, GC content, gene length, average intron length, “minimal intron” density, and recombination rate in the chicken genome (Figure 6). A correlation between recombination rate and GC content in the chicken genome has been recently reported [23], and these authors therefore link recombination rate with the transcriptome map, as reported in the current study. This can be explained by the more open chromatin structure of the transcriptionally active RIDGEs, which would also facilitate recombination within these regions. Furthermore, “minimal introns” have been reported to be GC-rich and to enhance the rate at which mRNA is exported from the cell nucleus [30] (Yu et al. 2002). These findings link the “minimal introns” distribution via GC content with the transcriptome map in the current study. This can be explained, at least in part, by the need for efficient export of highly expressed mRNA from the nucleus. Many genomic characteristics in eukaryotic genomes, such as RIDGEs, early replication and recombination, appear to be linked. RIDGEs are associated with higher expression, higher gene density, higher GC content, shorter gene introns, shorter genes, higher “minimal intron” density, and higher recombination rate (Figure 6). This is congruent in human studies, in which similar correlations were found [9]. Shorter introns and shorter genes in RIDGEs may indicate the need for increased transcription efficiency. Castillo-Davis et al. [31] showed that introns in highly expressed genes are substantially shorter than those in genes that are expressed at low levels in the human genome, and the authors hypothesized that transcription efficiency is enhanced when intron length is shorter. The clustering of highly expressed genes in RIDGEs therefore would result in clustering of genes with, on average, shorter introns. Although GC content, gene density, gene length, average intron length, “minimal intron” distribution and recombination rate are all correlated with gene transcriptional activity in the chicken genome, the exact causative mechanisms of these relationships are still unknown.

RIDGEs are relatively conserved between chicken and human

In comparing evolutionary break points between RIDGEs and anti-RIDGEs, we found a higher number of break points within RIDGEs than anti-RIDGEs in both the chicken and the human genome. Similar as for recombination, it is possible that the more open chromatin structure within RIDGEs facilitates an increase in the likelihood of rearrangement events, and thus in an increase in the observed syntenic breaks.

Although RIDGEs clearly show an increase in the number of evolutionary break points, we also showed a significantly higher number of RIDGE-to-RIDGE break points between the chicken and human genomes. Hence, although RIDGEs are more prone to be interrupted by evolutionary break points, there still seems to be an evolutionary constraint that favors recombination between RIDGEs, i.e. the resulting parts of a “broken RIDGEs” from one species were more likely to stay together with a part of another broken RIDGE during genome evolution, thereby keeping specific genes together within RIDGEs. In other words genes within a RIDGE in one species are likely to end up in a RIDGE in another species even when syntenic rearrangements occur. There are in total 11,407 1-to-1 human-chicken homolog genes downloaded via biomaRt [32]. Of these genes, 1,351 are located In RIDGEs and 857 genes are located in anti-RIDGEs in the human genome. 27% of these 1-to-1 human-chicken homolog genes (361 out of 1351 genes) located in human RIDGEs are also located in chicken RIDGEs (p-value smaller than 2.2 x 10-16, Chi-square tests). This again supports our hypothesis that genes within a RIDGE in one species are likely to end up in a RIDGE in another species.

This result suggests that the clustering of specific genes is not so much important, but rather the clustering of any genes that are highly expressed. The relative low number of syntenic breaks within anti-RIDGEs, on the other hand, might be linked to another feature of vertebrate chromosomes, namely the occurrence of regions with a relatively low number of genes, so called “gene deserts” [33]. In particular, the so-called “stable gene deserts” co-localize with developmentally active genes and genes coding for transcription factors, both gene types that generally show relatively low levels of expression. These “stable gene deserts” showed extremely low numbers of syntenic breaks [33].

Our results clearly show the existence of a higher level organization of the vertebrate genome affecting not only the expression of genes but also other features such as recombination and genome rearrangements during evolution.

Conclusion

This is the first study describing a transcriptome map in birds. This study has revealed regional regulation of gene expression in chicken that is consistent with previous studies in flies and mammals [8, 10, and 12]. Since features correlating with high regional transcription are more pronounced in the microchromosomes leading to overall higher expression compared to genes on the macrochromosomes. Our analysis on evolutionary break points shows that the regional regulation of gene transcription is relatively conserved between chicken and human. Given the evolutionary position of chicken on the phylogenetic tree, our results provide a unique perspective for future comparative studies on transcriptome maps between vertebrate species.

Methods

Gene expression data

The gene expression data used in this analysis was obtained from a gene expression survey in chicken brain, bursa of Fabricius, kidney, liver, lung, small intestine, spleen and thymus, using the chicken 20k oligonucleotide microarray (see below). Five biological replicates were used for each tissue type, resulting in a total of 40 arrays. Each individual sample was compared to the pooled reference, and data was normalized using the R [34] package limma [35]. The mean expression value for each Ensembl gene was calculated for each tissue type, and the average expression value of each Ensembl gene was determined by calculating the median expression values across all eight tissues.

The microarray data have been deposited in the Gene Expression Omnibus (GEO) public repository [24]. The accession number for the series is GSE17108, and the sample series can be retrieved with accession numbers from GSM427873 to GSM427912. The sample series contains the raw data (median signal) of each Cy5 (red) and Cy3 (green) channels as well as the normalized data for each microarray.

Chicken 20k array platform and oligonucleotide probe re-annotation

The chicken 20k array was obtained from ARK-Genomics [36]. The array design has been published in Gene Expression Omnibus with the platform name GPL8861 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=tjwjpscyceqawjk&acc=GPL8861).

The probe sequences of the chicken 20k oligonucleotide microarray used in this study were designed based on chicken genome assembly WASHUC1 (Dec. 2004), and all sequences were mapped to the chicken genome assembly WASHUC2. An updated array re-annotation file based on Ensembl 50 is available at EADGENE Oligo Set Annotation Files homepage [37]. Of the total 20,460 oligonucleotide probes on the chicken 20k array, 13,431 mapped to unique locations in the chicken genome. All the probes for genes that mapped to chromosome “unknown” were excluded in the analysis, and all probes for genes on chromosome 16, 25, and W were excluded due to the very low number of probes that mapped to those chromosomes. For probes that mapped to the same known Ensembl gene ID [38], the expression data were averaged and assigned to the Ensembl gene. In total, in this study, 12,983 oligo probes were used that mapped to 11,361 unique chicken Ensembl gene IDs located on 27 chromosomes.

Identification of RIDGEs in the chicken genome

Individual gene expression data was ordered according to the middle position of the gene. A Robust Scatter Plot Smoothing (function *runmed* in R package stats) technique was applied to each chromosome separately, with a window size of 39 genes, i.e. the expression value of each gene was replaced by the median expression value of the neighboring 39 genes. Similar to the definition for RIDGEs in humans [8], here we defined a RIDGE by window size for calculating median expression, minimum length of the run, and the threshold for the lower limit of the median. The selection of window size of 39 genes was based on the following two points: 1) Permutation analysis performed by both Caron et al. [8] and our analysis indicated a window size of 39 genes gives a reasonable number of RIDGEs; 2) To be able to compare the results of RIDGE identification between human and chicken, we decided to use the same threshold as described by Caron et al. The bigger the window size is, the smaller number of RIDGEs will be identified as indicated in the permutation results in Additional file 3.

The threshold for RIDGEs was set to 1.19 times the genomic median value (the data are log2 transformed, and the values used here is the running median values of a window size of 39 genes) along the length of a run of at least 10 median values. The threshold used for anti-RIDGEs was a median expression of 0.78 times the genomic median. The thresholds used for the classification of the RIDGEs and anti-RIDGEs were chosen such that RIDGEs and anti-RIDGEs each cover 10% of the genome.

Correlation analysis between tissue-specific transcriptome maps

Spearman rank correlation test was performed to test for pairwise correlations among the transcriptome maps on all the chromosomes (applied to the running median with window size of 39 genes). The running median expression values are not normally distributed, and the non-parametric Spearman correlation test was used on the ranks of the paired transcriptome maps.

Random permutation tests for RIDGE identification in chicken

Random permutation tests were done in R by permuting the genomic locations of Ensembl genes and repeating the RIDGE analysis 10,000 times to create 10,000 random transcriptome maps. The number of RIDGEs identified in these 10,000 random transcriptome maps was compared to the actual number of identified RIDGEs in this analysis using the same threshold.

Syntenic break points

Human-chicken synteny block data from Bourque et al. [25] was used in this study, and genomic locations of synteny blocks from assembly WASHUC1 (Dec 2004) were mapped to assembly WASHUC2 (May 2006) using BLAT (see Additional file 4). Each end of every syntenic block was considered a break point, and the number of break points in RIDGEs and anti-RIDGEs was subsequently summarized.

Human gene expression data

Human Transcriptome Map data was downloaded from the HTM website [26]. We selected Affymetrix U133A human whole genome array data from seven tissues (thymus, spleen, lung, small intestine, brain, liver, and kidney) from a healthy individual; data (normalized data) was log2 transformed and the median expression value across the seven different tissues was used to build the transcriptome map. RIDGEs and anti-RIDGEs were identified using the same approach as for the chicken data.

Genome characteristics of RIDGEs and anti-RIDGEs in chicken

Genomic location, transcript length, exon number and GC content for the individual Ensembl chicken genes were downloaded from the Ensembl genome database using biomaRt [32]. The averaged intron length was calculated by averaging the intron length of all transcripts per gene. The statistical test for differences in average intron length, gene length, gene density, and GC content between RIDGEs and anti-RIDGEs was performed using Wilcoxon rank-sum test (function Wilcox.test function in R package stats).

GO term enrichment analysis

GO term enrichment analysis was performed using R package Gostats [28]. The conditional algorithm was used for the hypergeometric test. The gene annotation package for the GOstats analysis was built using R package AnnotationDbi [39]. Mapping of chicken Ensembl gene IDs and other genomic information (e.g. entrezgene) was performed using the R package biomaRt [32].

List of abbreviations used

MHC: Major Histocompatibility Complex; TF: Transcription Factor; SAGE: Serial Analysis of Gene Expression; RIDGE: Regions of Increased Gene Expression; GFP: Green Fluorescent Protein; EST: expressed sequence tag; GO: Gene Ontology.

Authors' contribution

HN carried out the experiment, performed data analysis and drafted the manuscript, JB helped with statistical analysis of this work, HJM, RC, and MG helped with interpretation of the results, all authors were involved in improving the manuscript. The final version of the manuscript was approved by all the authors.

Acknowledgements

This work was financed by European Union grant FOOD-CT-2004—506416 (EADGENE). We would like to acknowledge Aart Lammers and Basav Hangalapura from Wageningen University for helping with collection of chicken tissues, Richard Talbot from Ark-Genomics for providing the microarrays, Pieter Neerincx and Jack Leunissen from Wageningen University and Christophe Klopp and Pierrot Casel from INRA for help with updating the annotation for the chicken 20k oligo array.

References

1. Kielman MF, Smits R, Devi TS, Fodde R, Bernini LF: **Homology of a 130-kb region enclosing the alpha-globin gene cluster, the alpha-locus controlling region, and two non-globin genes in human and mouse**. *Mamm Genome* 1993, **4**(6):314-323.
2. The MHC sequencing consortium: **Complete sequence and gene map of a human major histocompatibility complex**. *Nature* 1999, **401**(6756):921-923.
3. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH: **Zebrafish hox clusters and vertebrate genome evolution**. *Science* 1998, **282**(5394):1711-1714.

4. Garcia-Fernandez J: **The genesis and evolution of homeobox gene clusters**. *Nat Rev Genet* 2005, **6**(12):881-892.
5. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order**. *Nat Rev Genet* 2004, **5**(4):299-310.
6. Sproul D, Gilbert N, Bickmore WA: **The role of chromatin structure in regulating the expression of clustered genes**. *Nat Rev Genet* 2005, **6**(10):775-781.
7. Michalak P: **Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes**. *Genomics* 2008, **91**(3):243-248.
8. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains**. *Science* 2001, **291**(5507):1289-1292.
9. Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH: **The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes**. *Genome Res* 2003, **13**(9):1998-2004.
10. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the Drosophila genome**. *J Biol* 2002, **1**(1):5.
11. Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI: **Large clusters of co-expressed genes in the Drosophila genome**. *Nature* 2002, **420**(6916):666-669.
12. Mijalski T, Harder A, Halder T, Kersten M, Horsch M, Strom TM, Liebscher HV, Lottspeich F, de Angelis MH, Beckers J: **Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues**. *Proc Natl Acad Sci U S A* 2005, **102**(24):8621-8626.
13. Singer GA, Lloyd AT, Huminiecki LB, Wolfe KH: **Clusters of co-expressed genes in mammalian genomes are conserved by natural selection**. *Mol Biol Evol* 2005, **22**(3):767-775.
14. Gierman HJ, Indemans MH, Koster J, Goetze S, Seppen J, Geerts D, van Driel R, Versteeg R: **Domain-wide regulation of gene expression in the human genome**. *Genome Res* 2007, **17**(9):1286-1295.
15. Chakalova L, Debrand E, Mitchell JA, Osborne CS, Fraser P: **Replication and transcription: shaping the landscape of the genome**. *Nat Rev Genet* 2005, **6**(9):669-677.
16. Croft JA, Bridger JM, Boyle S, Perry P, Teague P, Bickmore WA: **Differences in the localization and morphology of chromosomes in the human nucleus**. *J Cell Biol* 1999, **145**(6):1119-1131.

17. Habermann FA, Cremer M, Walter J, Kreth G, von Hase J, Bauer K, Wienberg J, Cremer C, Cremer T, Solovei I: **Arrangements of macro- and microchromosomes in chicken cells**. *Chromosome Res* 2001, **9**(7):569-584.
18. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.
19. Rodionov AV: **Micro vs. macro: structural-functional organization of avian micro- and macrochromosomes**. *Genetika* 1996, **32**(5):597-608.
20. Morescalchi A, Odierna G, Olmo E: **Karyological relationships between the Cyptobranchid salamanders**. *Specialia* 1977, **15**:1579.
21. Morescalchi A, Odierna G, Olmo E: **Karyology of the primitive salamanders, family Hynobiidae**. *Experientia* 1979, **35**:1434-1436.
22. Mengden GA, Stock AD: **Chromosomal evolution in Serpentes: a comparison of G and C chromosome banding patterns of some Colubrid and Boid genera**. *Chromosoma* 1980, **79**:53-64.
23. Groenen MA, Wahlberg P, Foglio M, Cheng HH, Megens HJ, Crooijmans RP, Besnier F, Lathrop M, Muir WM, Wong GK, 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**(3):510-519.
24. **National Center for Biotechnology Information (NCBI) Gene Expression Omnibus**: [http://www.ncbi.nlm.nih.gov/geo/]
25. Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G: **Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages**. *Genome Res* 2005, **15**(1):98-110.
26. **Human Transcriptome Map website**: [http://bioinfo.amc.uva.nl/HTMseq/]
27. **Gene Ontology website**: [http://www.geneontology.org/]
28. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association**. *Bioinformatics* 2007, **23**(2):257-258.
29. Goetze S, Mateos-Langerak J, Gierman HJ, de Leeuw W, Giromus O, Indemans MH, Koster J, Ondrej V, Versteeg R, van Driel R: **The three-dimensional structure of human interphase chromosomes is related to the transcriptome map**. *Mol Cell Biol* 2007, **27**(12):4475-4487.
30. Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GK: **Minimal introns are not "junk"**. *Genome Res* 2002, **12**(8):1185-1189.
31. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: **Selection for short introns in highly expressed genes**. *Nat Genet* 2002, **31**(4):415-418.

32. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis**. *Bioinformatics* 2005, **21**(16):3439-3440.

33. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L: **Evolution and functional classification of vertebrate gene deserts**. *Genome Res* 2005, **15**(1):137-145.

34. R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

35. Smyth GK: **Limma: linear models for microarray data**. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York: Springer; 2005:397-420.

36. **Ark-Genomics, Roslin institute, UK** [http://www.ark-genomics.org/]

37. **EADGENE Oligo Set Annotation Files homepage** [http://www.eadgene.info/TheProject/Integration/BiologicalresourcesandfacilitiesWP1/EADGENEOligoSetsAnnotationFiles/tabid/324/Default.aspx]

38. **Ensembl Genome Database** [http://www.ensembl.org/]

39. Pages H, Carlson M, Falcon S and Li N: **AnnotationDbi: Annotation Database Interface**. 2008, R package version 1.4.0.

Additional data files

Additional file 1

Title: Genomic location of RIDGEs and anti-RIDGEs.

Description: Genomic location of RIDGEs and anti-RIDGEs identified in the chicken genome in this study.

http://www.biomedcentral.com/imedia/3710740852964028/supp1.xls

Additional file 2

Title: Correlations of transcriptome maps in different tissues.

Description: All pairwise correlations between the tissue-specific transcriptome maps.

http://www.biomedcentral.com/imedia/5646842942964028/supp2.xls

Additional file 3

Title: Random permutation test.

Description: Random permutation test results for RIDGE identification with different window sizes.

http://www.biomedcentral.com/imedia/5561063582964028/supp3.xls

Additional file 4

Title: Positions of the synteny block in the chicken genome.

Description: Genomic positions of the ends of the synteny block on genome build WASHUC2.

http://www.biomedcentral.com/imedia/1442559527296402/supp4.xls

Additional file 5

Title: Evolutionary breaks within RIDGEs and anti-RIDGEs.

Description: Chi-square test of evolutionary break points within RIDGEs and anti-RIDGEs.

http://www.biomedcentral.com/imedia/1592715411296402/supp5.xls

Additional file 6

Title: GO enrichment analysis for genes in RIDGEs and anti-RIDGEs.

Description: Enriched GOBP terms for all genes located within RIDGEs and anti-RIDGEs. BY: adjusted p-values for the Benjamini & Yekutieli step-up FDR controlling procedure.

http://www.biomedcentral.com/imedia/2980093296402834/supp6.xls

Chapter 7

General discussion

The completeness of the chicken genome sequence in 2004 [1] represented a landmark in chicken biology and has opened new possibilities to increase our understanding of the biological functions of the genes within the chicken genome. As introduced in Chapter 1, the chicken sequence also provides a valuable reference for investigating the evolution of more general mechanisms of gene transcription in vertebrate genomes. An important challenge in the post-sequence era of chicken biology is determining the functional role of known genes and identifying previous un-characterized genes. In this thesis I have described two genome-wide gene expression surveys (Chapter 4 and 5) across adult chicken tissues and embryonic stages/tissues and used these to characterize the general expression profiles of genes in the chicken genome. These resources proved to be a valuable resource to understand basic mechanisms of gene regulation in vertebrates (chapter 6) and also in the future will further help to improve the accuracy of gene annotation in the chicken and for further studies to investigate gene transcription regulation and evolution in vertebrates (Chapter 4).

7.1 The gene models in the chicken genome

Microarrays can only monitor expression of genes which are included on the array while the problems of probe annotation on the array, in particular for farm animal species, has been introduced in Chapter 2.

The chicken 20K oligoarray used in this project was originally designed based on the first chicken assembly (WASHUC1, Mar 2004) and mainly based on the gene models of Ensembl release version 30 [2]. This platform includes 20,460 probes targeting 14,748 unique Ensembl genes and other expressed sequences (e.g., EST, cDNA clones). The second chicken assembly (WASHUC2, May 2006) was released with higher sequence quality and coverage, requiring an update of the annotation of the probes on the array and to estimate the probe specificity using the updated information. In total, 14,900 probes (out of 20,460 probes) targeting 8,792 Ensembl genes were uniquely mapped to the second chicken assembly using oligoRAP [3]. The 5,956 Ensembl genes that were missed was mainly due to the following reasons: 1) the update of the assembly from WASHUC1 to WASHUC2 resulted in some changes of sequences; 2) gene models in Ensembl from release 30 were updated in the current version and a relatively large number of gene models from previous assembly was updated or removed; 3) the stringent settings of oligoRAP to find hits and our decision to exclude probes that have more than one perfect hit in the genome. The latter to increase the accuracy of the functional annotation of the probes on the array in order to be able to unequivocally interpret the expression patterns obtained. Furthermore, many probes (from

14,900 mapped probes) with unique perfect hits in the WASHUC2 assembly and e.g. showing a tissue-specific expression profile do not map to known Ensembl gene models. The individual examples of brain-specific probes (Chapter 4) imply that the current prediction of the 3' UTR of chicken genes is not perfect in the WASHUC2 genome assembly. In addition other expressed probes not mapped to known genes imply that the chicken genome contains a large number of still un-annotated transcribed regions.

In the current chicken genome assembly (WASHUC2, May 2006), still many sequences have not been assigned correctly to a known chromosome (chr_random sequences), and the 10 smallest microchromosomes are still missing. Although a significant proportion of the chicken genome could not be assigned correctly to the current assembly or is even completely missing, new sequencing technologies are expected to further improve future genome assemblies of the chicken genome. The re-sequenced chicken genome at Washington University using 454 sequencing technology (Roche) and the new assembly will provide a better reference for microarray probe mapping, and therefore, provide more accurate probe function annotation on the chicken 20K oligoarray platform used in this study. This shows that re-annotation of the probes on the array using tools like OligoRAP is needed for every new genome build. Furthermore, this also shows the need for increased efforts of (manual) annotation of chicken genes. Such efforts will further increase the usefulness of the resources described in this these in the future. Our expression data of the 5,560 un-mapped probes (on the chicken 20K oligoarray) in the two expression surveys across different tissues and developmental stages (Chapter 4 and 5) provides further evidences for the expression profiles of many of the un-characterized transcribed regions (both new genes as well as unknown alternative splicing variants or known genes) in the chicken genome. This information will further help to improve the much needed further functional annotation of the chicken genome.

In addition to a better genome assembly, annotation pipelines like OligoRAP will need to be updated too to adapt the annotation strategies to our changing insights in gene expression. By doing so, we will ensure the availability of the most accurate probe annotation available to study gene expression using microarrays.

7.2 Compactness of housekeeping genes

As described in chapter 4, housekeeping genes, compared to tissue-specific genes, are relatively compact, i.e. shorter gene, shorter coding sequence length, shorter average intron size, and shorter intergenic region. This suggests selective constraint of compactness on housekeeping gene (widely expressed genes). The GO enrichment analyses show that these "housekeeping genes" are involved in essential biological processes. This finding was further validated in Chapter 5 using gene expression data surveying completely different stages during chicken development. As discussed in Chapter 5, about 81% of "housekeeping genes" in adult tissues were also identified being "housekeeping genes" in embryonic stages/tissues. The large overlap of the two groups of housekeeping genes identified at two distinct developmental stages (adult and embryonic stages) confirms the housekeeping functions of most of these identified "housekeeping genes" in both analyses. The compactness of housekeeping genes in both analyses (Chapter 4 and 5) suggests a selection for compactness on housekeeping genes by reducing the cost of transcription.

In contrast, tissue-specific genes are less compact and have larger f non-coding (NC) sequences (introns and intergenic regions). Active regulatory elements (REs) from anonymous NC sequences have been identified comparing human and draft zebrafish genomes, and were reported to be strongly involved in modulating tissue-specific expression of a green fluorescent protein reporter vectors using zebrafish transient transgenesis [5]. A similar finding was also reported in Arabidopsis where a small intergenic region was found to drive exclusive tissue-specific expression of the adjacent genes [6]. Therefore, the larger NC regions of tissue-specific genes found in this thesis may suggest that the regulation of expression of these genes in a number of specific tissues might have resulted in more complex regulation of transcription. A larger number of cis-regulatory elements might be involved in tissue-specific gene transcription and this would need larger regulatory "spaces" resulting in larger introns and intergenic regions in these genes.

7.3 Gene expression conservation in vertebrates

The expression of orthologous genes is generally well conserved as compared to random gene pairs (Chapter 4). The results described in this thesis suggest that gene expression is under some selection constraint during evolution. However, the gene expression conservation study as described in Chapter 4 still has a number of limitations. First of all, different tissue samples used in different gene expression surveys are mixtures of cells of different types within certain tissues. For example, the majority of the tissues from the different organs also include general cell types such as those involved in the formation of blood vessels and connective tissues. The gene expression levels measured in the surveys therefore included in this thesis are only a crude estimate of the average expression level in the different tissues analysed. Secondly, for the gene expression conservation study, although the sampled tissues in the different species were all from adult individuals, the ages may not be directly comparable across these species. The term "adult" only implies a crude estimate of the time point during the development of the individuals in the different species. Thirdly, an obvious limitation in combining data from several different species (as well as different microarray platforms) is that as more species are included, fewer representative genes are found to be common amongst all.

Although there are limitations as described above, I have shown in Chapter 4 that the gene expression pattern of orthologous gene pairs, compared to random gene pairs, are more conserved. This is in agreement with the results obtained in a comparison of different mammals [7, 8]. In our study we extended these findings to a wide range of vertebrates including mammals, birds, and amphibians. Although the number of 1:1:1 orthologous genes among the three species was limited, the conserved gene expression patterns of these 1:1:1 orthologous genes suggest that gene expression is under selection constraint in vertebrates during evolution. The finding on orthologous gene expression conservation in Chapter 4 has extended the range of species for gene expression conservation studies from mammals to birds and amphibians. By comparing distant species, our results provide evidence for gene expression conservation within vertebrates rather than only in mammals.

Furthermore, we show similarities of homologous tissues in terms of expression, brain tissues are highly correlated within the three species (mouse, chicken, and frog) indicating that the stronger evolutionary constrains posed on brain. In contrast, intestine and kidney show relatively low conservations. Kidneys have diverged functions in different vertebrate species [9] and intestines subject to greater environmental influence, genes expressed in these two tissues may be more likely to take on new roles of diverge in expression as means of adaptation.

7.4 Regional regulation of gene transcription

Chapter 6 describes the first study constructing a transcriptome map in birds and confirms the existence of regional differences on transcription regulation in the chicken genome. The results reveal the presence of two distinct types of chromosomal regions characterized by clusters of highly or lowly expressed genes. Regions with clusters of highly expressed genes have higher gene densities, shorter genes, shorter average intron and higher GC content compared to regions with clusters of lowly expressed genes. Furthermore, the housekeeping genes are in favor of being located in RIDGEs in the chicken genome as discussed in Chapter 4, this indicates that these genes need to be expressed at relative higher levels and at a larger number of physiological conditions.

In vertebrates, transcription of protein-coding genes is performed by RNA polymerase II. Genes transcribed by RNA polymerase II typically contain two distinct families of cis-acting transcriptional regulatory DNA elements: (a) a promoter, which is composed of a core promoter and nearby (proximal) regulatory elements, and (b) distal regulatory elements, which can be enhancers, silencers, insulators, or locus control regions (LCR) [10]. The findings in Chapter 6 suggest the existence of multi-level gene regulation: transcription factors (bind to promoter regions) determine whether a gene will be expressed and also establish a basic level of transcription; in addition, there is a substantial effect of the region where the gene is positioned. Furthermore, it was shown that large intergenic regions lacking transcribed genes and classified as gene deserts, may play a role in the regulation of neighboring genes [11]. Again, these findings clearly show the complexity of the regulation of gene transcription in vertebrate genomes.

The regional regulation of transcription has been reported to be relatively conserved between the mouse and human genome [12]. , Our comparative analysis between the chicken and human transcriptome maps (Chapter 6 of this thesis) suggests that the regions with clusters of highly expressed genes are relatively conserved between the two genomes as well. Given the evolutionary position of chicken on the phylogenetic tree, our results clearly show that the regional regulation is a common mechanism regulating gene expression in vertebrate species. The exact mechanism underlying this regional regulation of transcription in genomes is still largely unknown, but the conservation of such mechanism among human, mouse, and chicken [11, Chapter 6 of this thesis] clearly shows that it is under strict evolutionary constraints to maintain normal biological functions in vertebrate genomes.

The regional regulation of transcription could be regulated either through an activating or suppressive mechanism (RIDGEs and anti-RIDGEs) or both. Gene activation or suppression often is accompanied by changes in the histone code and/or DNA methylation [13]. It is not known whether the histone codes also play a role in the regional regulation of gene expression reported in this thesis, but histone modification can spread over large genomic distances and have been reported to be associated with activating gene expression [14, 15]. The ability to perform genome-wide analysis of histone modifications will enable us to identify regional effects of histone modifications on gene expressions, this will help us to understand to what extent histone modifications are involved in regional regulation of gene expression in the genome described in this thesis.

7.5 Gene expression study in chicken in the future

The evolving knowledge of eukaryotic transcriptomes has shown that the eukaryotic transcriptome is much more complex than previously anticipated, involving overlapping transcripts, transcribed intergenic regions and abundant non-coding RNAs [16]. Expression microarrays are currently the most widely used methodology for transcriptome analysis, although some limitations persist. These include hybridization and cross-hybridization artifacts, dye-based detection issues and design constraints that preclude or seriously limit the detection of RNA splice patterns and previously unmapped genes [17]. A new method, called RNA-Seq [17], which uses high-throughput direct sequencing of the transcripts within a specific sample, can provide a more comprehensive understanding of this complexity of the transcriptome. RNA-Seq involves direct sequencing of cDNAs using high-throughput sequencing technologies, thereby allowing the level of transcription from a particular genomic region to be quantified from the density of corresponding reads. Unlike array-based approaches, RNA-Seq gives a potentially comprehensive view of the transcriptome, and avoids the bias of only focusing on previously identified transcripts. Another advantage is its ability to provide information on transcripts that are expressed at very low levels, limited only by the total number of reads that are generated [17]. A recent study surveying the human transcriptome using RNA-Seq showed that, based on known transcripts, RNAseq can detect 25% more genes than microarrays and exon skipping was found to be the most prevalent form of alternative splicing [18].

Furthermore, another recent study [19] reported that the deep sequencing used in RNA-Seq experiments provides a major advantage in robustness, comparability and richness of expression profiling data and is expected to boost collaborative, comparative and integrative genomics studies among different experiments. The real challenge for microarrays in the coming years will be to remain up to date. Our understanding of the transcriptome is constantly evolving, and this makes it difficult for microarrays to stay current.

In this thesis, both expression surveys across tissues were performed using a chicken 20K oligoarray. Using this array more than half of the chicken genes in terms of Ensembl genes (8792 out of 15,908 known protein-coding genes) have been surveyed to study the regulation of gene transcription in the chicken. The limited number of genes included in the analyses described in this thesis was mainly due to the restrictions of this array platform as well as the still limited available annotation of the chicken genome. In the near future, the study of genome-wide gene expression will probably shift to sequencing-based technology because of the described advantages of the new technology. This will not only result in a more unbiased view of the transcriptome, but more importantly, it will boost further annotation of the chicken genome. In parallel, new developments in next generation sequencing will further improve the current genome assembly of the chicken, ultimately providing a more comprehensive view of this birds genome including the genes located on the currently still missing micro-chromosomes. .

7.6 Conclusions and future perspectives

The ultimate goal of genome research in chicken is the discovery of genes and regulatory regions and to understand the biological functions of these genes and their related regulatory networks. This knowledge can lead to a better understanding of candidate genes that perform key roles under specific experimental conditions. The research presented in this thesis resulted in the development of genome-wide gene expression resources for the chicken research community and these resources should provide a global picture of gene expression for other researchers in chicken biology, developmental biology or related fields. A number of methods to analyze microarray data and to extract biological information have been described. Selection on economy for compactness of housekeeping genes was identified and discussed in chicken, and furthermore, a novel level of gene transcription regulation was discovered in birds and this mechanism was shown to be conserved between human and chicken.

Regarding the future in genome research in chicken, given the rapid developments of new genomic tools such as surveys of genome-wide CNV (copy number variations) and SNP (Single-nucleotide polymorphism) detection, together with genome-wide gene expression data, a global picture of the relative impact of CNV and SNP on gene expression can be studied. Integrating genomic data from different sources, rather than using gene expression data alone, will lead us to a better understanding of the mechanisms of gene expression regulation in the chicken.

Abbreviations

RE: regulatory element
NC: non-coding regions
CNV: copy number variations
SNP: Single-nucleotide polymorphism

References

1. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.
2. **Ensembl Genome Database**: [http://www.ensembl.org/]
3. Neerincx PB, Rauwerda H, Nie H, Groenen MA, Breit TM, Leunissen JA: **OligoRAP - an**

Oligo Re-Annotation Pipeline to improve annotation and estimate target specificity. *BMC proceedings* 2009, **3 Suppl 4**:S4.

4. **UCSC Genome Browser**: [http://genome.ucsc.edu/]
5. Shin JT, Priest JR, Ovcharenko I, Ronco A, Moore RK, Burns CG, MacRae CA: **Human-zebrafish non-coding conserved elements act in vivo to regulate transcription**. *Nucleic acids research* 2005, **33**(17):5437-5445.
6. Bondino HG, Valle EM: **A small intergenic region drives exclusive tissue-specific expression of the adjacent genes in Arabidopsis thaliana**. *BMC molecular biology* 2009, **10**:95.
7. Liao BY, Zhang J: **Evolutionary conservation of expression profiles between human and mouse orthologous genes**. *Molecular biology and evolution* 2006, **23**(3):530-540.
8. Khaitovich P, Enard W, Lachmann M, Paabo S: **Evolution of primate gene expression**. *Nature reviews* 2006, **7**(9):693-702.
9. Braun EJ: **Comparative renal function in reptiles, birds, and mammals**. *Seminars in Avian and Exotic Pet Medicine* 1998, **7**(2): 62-71.
10. Maston GA, Evans SK, Green MR: **Transcriptional regulatory elements in the human genome**. *Annu Rev Genomics Hum Genet* 2006, **7**:29-59.
11. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L: **Evolution and functional classification of vertebrate gene deserts**. *Genome research* 2005, **15**(1):137-145.
12. Singer GA, Lloyd AT, Huminiecki LB, Wolfe KH: **Clusters of co-expressed genes in mammalian genomes are conserved by natural selection**. *Mol Biol Evol* 2005, **22**(3):767-775.
13. Gierman HJ, Indemans MH, Koster J, Goetze S, Seppen J, Geerts D, van Driel R, Versteeg R: **Domain-wide regulation of gene expression in the human genome**. *Genome Res* 2007, **17**(9):1286-1295.
14. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR et al: **Genomic maps and comparative analysis of histone modifications in human and mouse**. *Cell* 2005, **120**(2):169-181.
15. Roh TY, Cuddapah S, Zhao K: **Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping**. *Genes Dev* 2005, **19**(5):542-552.
16. Louisa Flintoft: Transcriptomics: **Digging deep with RNA-Seq**. *Nature Reviews Genetics* 2008, **9**:568.
17. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying**

mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008, **5**(7):621-628.

18. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D et al: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome**. *Science* 2008, **321**(5891):956-960.
19. t Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT: **Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms**. *Nucleic acids research* 2008, **36**(21):e141.

Summary

The chicken (*Gallus gallus*) is an important model organism in genetics, developmental biology, immunology, evolutionary research, and agricultural science. The completeness of the draft chicken genome sequence provided new possibilities to study genomic changes during evolution by comparing the chicken genome to that of other species. The development of long oligonucleotide microarrays based on the genome sequence made it possible to survey genome-wide gene expression in chicken. This thesis describes two gene expression surveys across a range of healthy chicken tissues in both adult and embryonic stages. Specifically, we focus on the mechanisms of regulation of gene transcription and their evolution in the vertebrate genome.

Chapter 1 provides a brief history of the chicken as a model organism in biological and genomics research. In particular a brief overview is presented about expression profiling experiments, followed by an introduction to gene transcription regulation in general. Finally, the aim and outline of this thesis is presented.

An important aim of this thesis is to generate surveys of genome-wide gene expression data in chicken using microarrays. In **chapter 2**, we introduce microarray data normalization including background correction, within-array normalization and between-array normalization. Based on these results an analysis approach is recommended for the analysis of two-color microarray data as performed in the experiments described in this thesis. We also briefly explain the relevant methodology for the identification of differentially expressed genes and how to translate resulting gene lists into biological knowledge. Finally, specific issues related to updating microarray probe annotation in farm animals, is discussed. For the analysis of the microarray data in this thesis re-annotation of the probes on the chicken 20K oligoarray was done using the oligoRAP, analysis pipeline.

The vast amount of data generated from a single transcriptomics study makes it impossible to extract meaningful biological knowledge by manually going through individual genes from a list with hundreds and thousands of differentially expressed genes. In **chapter 3**, we present a practical approach using a collection of R/Bioconductor packages to extract biological knowledge from a microarray experiment in farm animals. Furthermore, a locally adaptive statistical procedure (LAP) analysis approach is used to identify differentially expressed chromosomal regions in a microarray experiment.

Chapter 4 presents a genome-wide gene expression survey across eight different tissues (brain, bursa of Fabricius, kidney, liver, lung, small intestine, spleen, and thymus from 10-week old chickens) in adult birds using a chicken 20K microarray. To a certain extent, most genes show some tissue-specific pattern of expression. Housekeeping and tissue-specific genes are identified based on gene expression patterns across the eight different tissues. The results show that housekeeping genes are more compact, i.e. are smaller, with shorter, coding sequence length, intron length, and smaller length of the intergenic regions. This observed compactness of housekeeping genes may be a result of selection on economy of transcription during evolution. Furthermore, a comparative analysis of gene expression among mouse, chicken, and frog showed that the expression patterns of orthologous genes are conserved during evolution between mammals, birds, and amphibians.

The chicken embryo has been a very popular model for developmental biology. To study the overall gene expression pattern in whole chicken embryos at different developmental stages and/or embryonic tissues, a genome-wide gene expression survey across different developmental and embryonic stages was performed (**chapter 5**). The study included four different developmental stages (HH stage 3, 10, 15, 22) and eight different embryonic tissues (brain, bursa of Fabricius, heart, kidney, liver, lung, small intestine, and spleen from HH stage 36). We were able to identify several embryonic stage- and tissue-specific genes in our analysis. Genomic features of genes widely expressed under these 12 conditions suggest that widely expressed genes are more compact than tissue-specific genes, confirming the findings described in chapter 4. The analysis of the differentially expressed genes during the different developmental stages of whole embryo indicates a gradual change in gene expression during embryo development. A comparison of the gene expression profiles between the same organs, of adults and embryos reveals both striking similarities as well as differences.

The overall goal of this thesis was to improve our understanding of the mechanisms of transcriptional regulation in the chicken. In **chapter 6**, a transcriptome map for all chicken chromosomes is presented based on the expression data described in chapter 4. The results reveal the presence of two distinct types of chromosomal regions characterized by clusters of highly or lowly expressed genes respectively. Furthermore, these regions show a high correlation with a number of genome characteristics, like gene density, gene length, intron length, and GC content. A comparative analysis between the chicken and human transcriptome maps suggests that the regions with clusters of highly expressed genes are relatively conserved between the two genomes. Our results revealed the presence of a higher order organization of the chicken genome that affects gene expression, confirming similar observations in other species.

Finally, in **chapter 7** I summarize the main findings and discuss some of the limitations of the analyses described in this thesis. I also discuss the different merits and shortcomings of studying gene expression using either microarrays or next-generation sequencing technology and propose directions for future research. The rapid developments in new-generation sequencing technology will facilitate better coverage and depth of the chicken genome. This will provide a better genome assembly and an improved genome annotation. The sequence-

based approaches for studying gene expression will reduce noise levels compared to hybridization-based approaches. Overall, next-generation sequencing is already providing greatly enhance tools to further improve our understanding of the chicken transcriptome and its regulation.

Samenvatting

De kip (*Gallus gallus*) is een belangrijk model organisme in genetica, ontwikkelings biologie, immunologie, evolutionair onderzoek en landbouwkundige wetenschappen. Het gereed komen van de eerste versie van de sequentie van het kippen genoom heeft nieuwe mogelijkheden gegenereerd om genomische veranderingen tijdens evolutie in kaart te brengen, door het kippen genoom te vergelijken met dat van andere soorten. De ontwikkeling van oligonucleotide microarrays gebaseerd op de genoom sequentie heeft het mogelijk gemaakt om genoom wijde gen expressie studies uit te voeren bij kip. Dit proefschrift beschrijft twee gen expressie studies gebruik makend van een aantal gezonde kippen weefsels in zowel volwassen en embryonale stadia. Specifiek richten wij ons op het regulatie mechanisme van gen transcriptie en hun evolutie in het vertebrate genoom.

Hoofdstuk 1 geeft een kort overzicht van de kip als model organisme in biologisch en genomisch onderzoek. Met name wordt een kort overzicht gepresenteerd over expressie profiling experimenten, gevolgd door een introductie van gen transcriptie regulatie in het algemeen. Tenslotte wordt het doel en de opbouw van dit proefschrift gepresenteerd.

Een belangrijk doel van dit proefschrift is om onderzoek te doen naar genoome wijde gen expressie in kip gebruikmakend van microarrays. In **hoofstuk 2** introduceren wij de normalizering van microarray gegevens, inclusief achtergrond correctie en normalisatie van zowel binnen als over arrays. Gebaseerd op deze resultaten wordt een analyse aanpak voorgesteld om de in dit proefschrift gegenereerde twee kleurige mircoarray data te analyseren. Verder leggen wij in het kort de relevante methodologie uit voor de identificatie van differentieel to expressie komende genen en hoe we deze lijsten met genen kunnen vertalen naar biologische kennis. Tenslotte is er specifieke aandacht voor het opwaarderen van de annotatie van microarray probes bij landbouwhuisdieren. Voor de analyse van de microarray data welke beschreven in dit proefschrift is de re-annotatie uitgevoerd van de 20K oligoarray probes met behulp van de analyse pijplijn oligoRAP.

Het merendeel van de data, welke gegenereerd is in een transcriptomics studie, maakt het onmogelijk om de hieruit betekenisvolle biologische kennis te extraheren door handmatig een lijst van duizenden differentieel tot expressie komende genen te bekijken. In **hoofdstuk 3** presenteren wij een praktische aanpak om biologische kennis uit een microarray experiment bij landbouwhuisdieren te halen, gebruikmakend van een verzameling softwareprogramma's binnen R/ Bioconductor. Verder is er een " locally adaptive statistical procedure" (LAP) analyse aanpak gebruikt om chromosomale gebieden met differentiële expressie in een microarray experiment op te sporen.

In **hoofdstuk 4** presenteren wij een genoom wijde expressie studie met 8 verschillende volwassen kippen weefsels (hersenen, bursa van Fabricius, nier, lever, long, dunne darm, milt en thymus elk van 10 weken oude kippen) gebruik makend van de 20K kippen

microarray. Tot op zekere hoogte laten de meeste genen een zekere mate van weefsel specifieke expressie patronen zien. De huishoud- en weefsel specifieke genen zijn geïdentificeerd op basis van de genexpressie patronen van de 8 verschillende weefsels. De resultaten geven aan dat de huishoudgenen compacter zijn, dat wil zeggen dat ze kleiner zijn, met kortere coderende sequentie, kortere intronlengte en een kleinere lengte van de gebieden tussen genen. De compactheid van de huishoudgenen kan een resultaat zijn van selectie op economische transcriptie tijdens evolutie. Verder laat een vergelijkende analyse van genexpressie tussen muis, kip en kikker zien dat de expressiepatronen van orthologe genen bewaard blijven tijdens evolutie tussen zoogdieren, vogels en amfibieën.

Het kippen embryo is een erg populair model systeem voor ontwikkelingsbiologie. Voor het bestuderen van het algemene genexpressie patroon in de embryo van de kip, van verschillende ontwikkelstadia en/of embryonale weefsels, wordt in **hoofdstuk 5** een genoom wijde genexpressie studie beschreven van verschillende ontwikkelings en embryonale stadia. Deze studie omvat vier verschillende ontwikkelingsstadia (HH stadium 3, 10, 15 en 22) en acht verschillende embryonale weefsels (hersenen, bursa van fabricius, hart, nier, lever, long, dunne darm en milt van HH stadium 36). Wij waren in staat om in onze analyse verschillende genen te identificeren voor de specifieke embryonale stadia en weefsels. Genomische kenmerken van de genen welke wijds tot expressie komen, in de twaalf onderzochte condities, compacter zijn dan de weefsel specifieke genen. Dit bevestigd de bevindingen welke beschreven zijn in hoofdstuk 4. De analyse van de genen welke differentieel tot expressie komen tijdens de verschillende ontwikkelingsstadia van de gehele embryo’s laat een graduele verandering zien in genexpressie tijdens embryonale ontwikkeling. Een vergelijking van genexpresie profielen tussen hetzelfde weefsel van volwassen en embryo laat zowel opvallende overeenkomsten als verschillen zien

Het doel van dit proefschrift was om onze kennis te verbeteren van het mechanisme van transcriptie regulatie van de kip. In **hoofdstuk 6** wordt een transcriptoom kaart van alle kippenchromosomen gepresenteerd, gebruik makend van de expressiegegevens beschreven in hoofdstuk 4. De resultaten laten de aanwezigheid zien van twee verschillende chromosomale regio’s die gekarakteriseerd worden door clusters van hoog en laag tot expressie komende genen. Bovendien laten deze gebieden een hoge correlatie zien met een aantal genoom specifieke kenmerken zoals gendichtheid, genlengte, intron lengte en GC gehalte. Een vergelijkende studie tussen de transcriptoom kaart van kip en mens met vergelijkbare weefsel types, suggereert dat de gebieden met clusters met genen welke hoog tot expressie komen relatief geconserveerd zijn tussen de twee genomen. Onze resultaten laten zien dat er een hogere orde organisatie van het genoom van de kip is die van invloed is op genexpressie, wat in overeenstemming is met vergelijkbare waarnemingen bij andere soorten.

Tenslotte worden in **hoofdstuk 7** de belangrijkste bevindingen nog eens samengevat en bespreek ik enkele beperkingen van de in dit proefschrift uitgevoerde analyses. Verder bediscuteer ik de voor- en nadelen van genexpressie studies waarbij gebruik gemaakt wordt van microarray of nieuwe generatie sequentie technologie. Daarnaast wordt een voorstel gedaan voor toekomstig onderzoek.

De snelle ontwikkeling van de nieuwe generatie sequentie technologie zal resulteren in een zowel een betere dekking als sequentiediepte van het kippengenoom. Dit levert op zijn beurt weer een betere genoom assembly op en een verbeterde genoom annotatie. Een op sequentie gebaseerde aanpak bij een genexpressie studie zal de achtergrond verminderen in vergelijking met de op hybridisatie gebaseerde benadering. Samenvattend, de nieuwe generatie sequentie technologie levert reeds sterk verbeterde gereedschappen om onze kennis van het kip transcriptoom en de regulatie daarvan verder te vergroten.

Acknowledgements

When the first time I arrived in Wageningen to start my MSc study in the summer of 2003, I could have never imagined that this small town would be my “home” in the following 6.5 years. During these years, I was very lucky to meet many people and made friends with many of them. In the last part of my thesis, which I am almost sure is the most read part of the entire thesis, I would like to express my profound gratitude to those who have helped me to find my way out during the last years in both scientific work and my personal life.

First of all I would like to thank my promoter Prof. Martien Groenen. Dear Martien, thank you for providing me the opportunity to start this PhD study in 2005, and for teaching me so many things (both scientific and non-scientific) and for being very supportive when I needed. Of course, also thank you for spending countless hours with me discussing research, proof reading papers and especially this dissertation, I really enjoyed my time working with you.

Second, I would like to thank my co-promoter Dr. Richard Crooijmans. Dear Richard, thank you for helping me with all the hard work taking samples in the beginning of the project for the experiments, and for providing very helpful assistance in the project during the last 4 years. Also thank you for being very understanding and supportive when I was experiencing difficult time during my PhD life. Of course, I also thank you for spending so much time with me discussing research, proof reading papers and this dissertation, and thank you for translating Samenvatting for this thesis.

Third, I would like to thank Prof. Johan van Arendonk. Dear Johan, thank you for all the encouraging words you have told me during our yearly R&O meetings in the past years, and thank you for being so supportive when things were not going very well in my PhD project. I could have more difficult time by the end of my PhD without your understanding and support. I really appreciated your kindness and help.

Then I would like to thank all my colleagues from ABGC, I have spent very nice time with you all, especially to Hinri Kerstens, I really enjoyed sharing the small “Area 51” with you in the last 4.5 years, and thank you for all the helpful tips/assistance you have provided for both my research and life, though there was only one down-side of sitting opposite your desk, you worked so hard and your PC was blowing warm airs all the time, that made it difficult for me to keep my hairstyle by the end of the day. Hendrik-Jan Megen, I really appreciated all the talk/discussion with you in the past years, you always inspired me and came up with brilliant ideas to improve my analysis, you were also one of people who encouraged me a lot when I felt depressed and lost during my PhD life. Jan van der Poel, thank you for being very patient when I was disturbing you with different questions about immunology, I have learned many things from our discussion, also thank you for sharing a lot of helpful tips about PhD and life which made feel a lot better when I was depressed. John Bastiaansen, thank you for your great help on statistics in my project, I really appreciated our collaboration which ended up as a nice paper. Albart Coster, thank you for all your help to improve my knowledge and skill with R, I learned a lot from you, and I also enjoyed teaching the R courses together with you, John, and Gerrit. Marleen Visker, thank you for helping with the Dutch summary. I also would like to thank my colleague PhD students from ABG, Andreia, Arun, Cecile, Martin, Luqman, Panya, Ghyslaine, Ane, Filippo and many others for the nice time we have spent together in ABG during our PhD lives.

I would like to thank our collaborators from Laboratory of Bioinformatics, Pieter Neerincx and Prof. Jack Leunissen, thank you for providing the assistances for microarray probe re-annotation, the collaboration also brought us to the EADGENE post-analysis workshop together in 2008 which ended up as four publications, and it was really a great team work. Pieter, I would like to specially thank you for helping me out with my work, and being very patient to explain me all the details, I have learned a lot from you about the probe re-annotation pipelines. I would like to thank Dr. Evert van Schothorst, and Prof. Jaap Keijer from Human and Animal Physiology for your very kindly help in the beginning of my project, I have gained my first knowledge from you about microarray normalization and downstream analysis when I was visiting your group for a couple of months in 2007. I would like to thank both Dr. Francesco Ferrari (University of Padova, Italy) and Dr. Silvio Bicciato (University of Modena and Reggio Emilia, Italy) for helping me with the analysis of differentially expressed chromosomal regions described in Chapter 3 of this thesis. I also would like to thank Paul Keizer and Prof. Fred van Eeuwijk from Biometris for the discussions about integrating multiple microarray datasets for a combined analysis in my project, although things worked out in a different way because of the limited time I had left in my PhD project, but I did learn quite some statistical knowledge from our discussion about microarray data integration from both of you.

Besides work, you always need friends wherever you are, here I would like to thank all of my friends for their friendships and help in the last years, especially to Gao Ran and Luo Jing, thank you both for the great friendships, and your care and supports when I was in difficult situations, I really appreciated everything you have done. Liu Huaifeng and Huang Huiyan, thank you both for your friendships, I really enjoyed the time we have spent together in the last few years. Anita and Paul, thank you for inviting me to your home several times in Wageningen and Almere, I really enjoyed being there with you, especially the days that I spent with Vidar. I would like to thank my friends, Guo Mingfei, Zhang Jin, Zhu Wei, Cai Jing, Xiao Yinghua, Wu Fei, Chen Xiaotong, Wang Fan, Lu Jing, Anord Kuzniar, Grazyna Truszkowska, Ma Lisong, Xu Fang, Lei Nan, Cheng Xu, Hao Xinning, Gao Xiao, Xing Ying, Wang Fengfeng, Wang Yiding, Lu Yu, Li Yuan, Wei Jianing, Zou Yuhua, Sun Yanyan, Tian Lijin, Yang Dongying and many others for your friendships and help. I won’t have enough space to mention every happy moment that I have experienced with every friend here in Wageningen, but I would like to thank you all for all the good time that you have brought to my life here, and I am very lucky to meet every one of you.

I also would like to thank all my neighbors in Hoevestein 3C, Christa, Andre, Carlijn, Djurre, Yvonne, Sylvia, Renske, Alicia, Joanna, I had a very good time sharing the same corridor with you all for the last several years (or few months).

Finally, my deepest gratitude to my parents for their endless love, support and encouragement, my sister for being there whenever I needed and for taking care of the family when I was far away from home. To them I dedicate this thesis. 最后，我要衷心的感谢爸爸妈妈无私的爱，鼓励，和支持，也要感谢姐姐在我远行的日子里，对家里的照料。我这些年最盼望的就是能回国和你们团聚，对于我来说，没有家人的团聚，未来的一切可能的所谓成就都将变得毫无意义。最后，我用童话里最常用的方式来结束致谢部分：期待回家，幸福快乐的生活在中国！☺

Haisheng Nie (聂海生)
Jan 2010
Wageningen, The Netherlands

Curriculum Vitae

Haisheng Nie (聂海生) was born on the 15th of July 1981 in SHAANXI (P.R. China), and he moved to Beijing with families and grew up there. After he finished his BSc in Animal Science at the China Agricultural University in Beijing in July 2003, he started MSc study in Wageningen University (The Netherlands) and majored in Animal Breeding and Genetics. From December 2004 through June 2005, he performed a minor thesis with Prof. Michel Georges in the University of Liege in Belgium. He graduated and received his Master degree in Animal Science and Aquaculture, specializing in Animal Breeding and Genetics in July 2005. In August 2005, he started his PhD project studying chicken transcriptomics at Wageningen University, within the Animal Breeding and Genomics Centre, which resulted in this thesis.

List of publications

Peer-reviewed publications

Nie H, Crooijmans RP, Bastiaansen JW, Megens HJ, and Groenen MA: Regional regulation of transcription in the chicken genome. ***BMC Genomics*** 2010, 11: 28.

Nie H, Neerincx PB, Poel J, Ferrari F, Bicciato S, Leunissen JA, Groenen MA: Microarray data mining using Bioconductor packages. ***BMC proceedings*** 2009, 3 Suppl 4:S9.

Neerincx PB, Rauwerda H, Nie H, Groenen MA, Breit TM, Leunissen JA: OligoRAP - an Oligo Re-Annotation Pipeline to improve annotation and estimate target specificity. ***BMC proceedings*** 2009, 3 Suppl 4:S4.

Neerincx PB, Casel P, Prickett D, Nie H, Watson M, Leunissen JA, Groenen MA, Klopp C: Comparison of three microarray probe annotation pipelines: differences in strategies and their effect on downstream analysis. ***BMC proceedings*** 2009, 3 Suppl 4:S1.

Hedegaard J, Arce C, Bicciato S, Bonnet A, Buitenhuis B, Collado-Romero M, Conley LN, Sancristobal M, Ferrari F, Garrido JJ, Groenen MA, Hornshoj H, Hulsegge I, Jiang L, Jimenez-Marin A, Kommadath A, Lagarrigue S, Leunissen JA, Liaubet L, Neerincx PB, Nie H, Poel J, Prickett D, Ramirez-Boo M, Rebel JM, Robert-Granie C, Skarman A, Smits MA, Sorensen P, Tosser-Klopp G, Watson M: Methods for interpreting lists of affected genes obtained in a DNA microarray experiment. ***BMC proceedings*** 2009, 3 Suppl 4:S5.

Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, Carta E, Dardano S, Dive M, Fasquelle C, Frennet JC, Hanset R, Hubin X, Jorgensen C, Karim L, Kent M, Harvey K, Pearce BR, Simon P, Tama N, Nie H, Vandeputte S, Lien S, Longeri M, Fredholm M, Harvey RJ, Georges M: Highly effective SNP-based association mapping and management of recessive defects in livestock. ***Nature genetics*** 2008, 40(4):449-454.

de Koning DJ, Jaffrezic F, Lund MS, Watson M, Channing C, Hulsegge I, Pool MH, Buitenhuis B, Hedegaard J, Hornshoj H, Jiang L, Sorensen P, Marot G, Delmas C, Le Cao KA, San Cristobal M, Baron MD, Malinverni R, Stella A, Brunner RM, Seyfert HM, Jensen K, Mouzaki D, Waddington D, Jimenez-Marin A, Perez-Alegre M, Perez-Reinado E, Closset R, Detilleux JC, Dovc P, Lavric M, Nie H, Janss L: The EADGENE Microarray Data Analysis Workshop (open access publication). ***Genet Sel Evol*** 2007, 39(6):621-631.

Jaffrezic F, de Koning DJ, Boettcher PJ, Bonnet A, Buitenhuis B, Closset R, Dejean S, Delmas C, Detilleux JC, Dovc P, Duval M, Foulley JL, Hedegaard J, Hornshoj H, Hulsegge I, Janss L, Jensen K, Jiang L, Lavric M, Le Cao KA, Lund MS, Malinverni R, Marot G, Nie H, Petzl W, Pool MH, Robert-Granie C, San Cristobal M, van Schothorst EM, Schuberth HJ, Sorensen P, Stella A, Tosser-Klopp G, Waddington D, Watson M, Yang W, Zerbe H, Seyfert HM: Analysis of the real EADGENE data set: comparison of methods and guidelines for data normalisation and selection of differentially expressed genes (open access publication). ***Genet Sel Evol*** 2007, 39(6):633-650.

Sorensen P, Bonnet A, Buitenhuis B, Closset R, Dejean S, Delmas C, Duval M, Glass L, Hedegaard J, Hornshoj H, Hulsegge I, Jaffrezic F, Jensen K, Jiang L, de Koning DJ, Le Cao KA, Nie H, Petzl W, Pool MH, Robert-Granie C, San Cristobal M, Lund MS, van Schothorst EM, Schuberth HJ, Seyfert HM, Tosser-Klopp G, Waddington D, Watson M, Yang W, Zerbe H: Analysis of the real EADGENE data set: multivariate approaches and post analysis (open access publication). ***Genet Sel Evol*** 2007, 39(6):651-668.

Abstracts in Conference proceedings

Nie H, Franssen-van Hal NL, van Schothorst EM, Crooijmans RP, and Groenen MA: A microarray survey of gene expression in healthy chicken tissues. Book of abstracts of the Genetica Retraite, 8-9 March 2007, Kerkrade, The Netherlands.

Nie H, Crooijmans RP, Bastiaansen JW, Megens HJ, and Groenen MA: Regional regulation of transcription in the chicken genome. Book of abstracts of the Biology of Genomes Meeting, 5-9 May 2009, Cold Spring Harbor, New York, USA.

Other publications

Ellen H Stolte, Haisheng Nie, Andy Cossins, Gert Flik, Huub FJ Savelkoul, BM Lidy Verburg-van Kemenade: Differential gene expression in the head kidney of common carp (Cyprinus carpio L.) following restraint stress or infection, revealed by transcriptome analysis. PhD Dissertation 2008, chapter 6: 109-124. (ISBN: 978-90-8585-199-8)

Publications in review or preparation

Nie H, Crooijmans RP, Lammers A, van Schothorst EM, Keijer J, Neerincx PB, Leunissen JA, Megen HJ, and Groenen MA: Gene expression in chicken reveals correlation with structural genomic features and conserved patterns of transcription in the terrestrial vertebrates. (Submitted)

Amaral A, Ferretti L, Megens HJ, Ramos-Onsins S, Crooijmans RP, Nie H, Pérez-Enciso M, Schook L, Groenen MA: Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing. (Submitted)

Nie H, Crooijmans RP, de Jong IM, de Bakker MA, Richardson MK, and Groenen MA: A Genome-wide Gene Expression Survey in Chicken embryos and embryonic tissues. (In preparation)

Training and Supervision Plan

Training and Supervision Plan

The Graduate School WIAS WAGENINGEN INSTITUTE of ANIMAL SCIENCES

The Basic Package (3 ECTS*)

WIAS Introduction Course (mandatory, 1.5 credits)	2006
Course on philosophy of science and/or ethics (mandatory, 1.5 credits)	2006

Scientific Exposure (15 ECTS)

International conferences

EADGENE Days, Oslo, Norway	2006
EADGENE-SABRE Days: "Genomics for Animal Health”, Utrecht, NL	2007
International Society of Animal Genetics, Amsterdam, NL	2008
4th EMBO Conference: From Functional Genomics..., Heidelberg, Germany	2008
Biology of Genomes Meeting, Cold Spring Harbor, New York, USA	2009

Seminars and workshops

How to participate in dynamic sciences, Arhnem, NL	2006
EADGENE Data Analysis Workshop,Tune, Denmark	2006
WIAS Science Day, Wageningen, NL (2x)	2008-09
EADGENE workshop on "post-analyses of microarray data" Lelystad, NL	2008
Genetics of milk quality, Wageningen, NL	2009
Genetica Retrait, Maastricht, NL (2x)	2006-07

Presentations (type of presentation)

Identification of transcription factor binding sites by sequence comparison and expression profiling, Oslo, Norway. (poster)	2006
A microarray survey of gene expression in chicken adult tissues. Genetica Retrait, Kerkrade, NL (oral)	2007
A microarray survey of gene expression in embryonic tissues WIAS Science Day, Wageningen, NL (oral)	2008
Array data mining using bioconductor packages. EADGENE workshop on "post-analyses of microarray data", Lelystad, NL (oral)	2008
Regional regulation of transcription in the chicken genome. The Biology of Genomes meeting, New York, USA (poster)	2009

In-Depth Studies (13 ECTS)

Disciplinary and interdisciplinary courses

Biology underpinning animal sciences: Broaden your Horizon, Wageningen, NL	2007
Mathematical modelling in biology, Wageningen, NL	2008

Advanced statistics courses

Microarray design and analysis course, Edinburgh, UK	2005
System biology course:"Statistical analysis of ~omics data", Wageningen, NL	2006
European Institute in Statistical Genetics, Liege, Belgium	2007
WIAS Advanced Statistics Course: Design of Animal Experiments, Wageningen, NL	2007
Statistical Analysis of Microarray Expression Data with R and Bioconductor, Copenhagen, Danmark	2007

Statutory Courses (3 ECTS)

The Course on Laboratory Animal Science, Utrecht	2006

Professional Skills Support Courses (3.5 ECTS)

Techniques for Writing and Presenting a Scientific Paper, Wageningen, NL	2006
French Language course 1st level, Wageningen, NL	2005
WIAS Workshop 'Career Coaching', Wageningen, NL	2007

Research Skills Training (5 ECTS)

Preparing own PhD research proposal	2005

Didactic Skills Training (14 ECTS)

Supervising 2 MSc theses	2006-07
Assisting practials MSc course Genomics	2006-08
Developing course exercies for MSc course Genomics	2008
Teaching PhD course: Introduction to R for Statistical analysis. April/Oct.(2x)	2008

Total Credits: 56.5

* One credit equals a study load of approximately 28 hours.

Colophon

This research is financially supported by the European Network of Excellence: EADGENE (European Animal Disease Genomics Network of Excellence for Animal health and Food Safety) (www.eadgene.info).

The printing of this thesis was funded by the Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands.

This thesis is printed by Wöhrmann Print Service, Zutphen, The Netherlands.