BIOINFORMATICS TOOLS FOR THE GENETIC DISSECTION OF COMPLEX TRAITS IN CHICKENS

Claudia Paola Cabrera Cárdenas

Doctor of Philosophy

The University of Edinburgh

School of Biological Sciences

2009

Abstract

This thesis explores the genetic characterization of the mechanisms underlying complex traits in chicken through the use and development of bioinformatics tools. The characterization of quantitative trait loci controlling complex traits has proven to be very challenging. This thesis comprises the study of experimental designs, annotation procedures and functional analyses. These represent some of the main 'bottlenecks' involved in the integration of QTLs with the biological interpretation of high-throughput technologies.

The thesis begins with an investigation of the bioinformatics tools and procedures available for genome research, briefly reviewing microarray technology and commonly applied experimental designs. A targeted experimental design based on the concept of genetical genomics is then presented and applied in order to study a known functional QTL responsible for chicken body weight. This approach contrasts the gene expression levels of two alternative QTL genotypes, hence narrowing the QTL-phenotype gap, and, giving a direct quantification of the link between the genotypes and the genetic responses. Potential candidate genes responsible for the chicken body weight QTL are identified by using the location of the genes, their expression and biological significance.

In order to deal with the multiple sources of information and exploit the data effectively, a systematic approach and a relational database were developed to improve the annotation of the probes of the ARK-Genomics *G. gallus* 13K v4.0 cDNA array utilized on the experiment. To follow up the investigation of the targeted genetical genomics study, a detailed functional analysis is performed on the dataset. The aim is to identify the downstream effects

through the identification of functional variation found in pathways, and secondly to achieve a further characterization of potential candidate genes by using comparative genomics and sequence analyses. Finally the investigation of the body weight QTL syntenic regions and their reported QTLs are presented.

Table of Contents

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Claudia P. Cabrera Cardenas)

Acknowledgements

First and foremost, I would like to thank my supervisors, DJ de Koning, Chris Haley, Andy Law and Sara Knott, for the great opportunity that they gave me. I am very grateful that I was allowed to work and learn from them. I would also like to express my deepest gratitude to DJ for all his support, advice, encouragement and patience throughout the PhD. I acknowledge financial support from Dorothy Hodgkins Awards and CONACYT.

I want to thank all NWE people. To the ones that were there when I arrived and gave me a great welcome (Elina K., Andreas K, Liz A.). Thanks to everybody for the shared sweets, smiles and unforgettable jokes and advice (Cecile, Suzanne, Georgia, Ariel, Ross, Ozzy, Dave W., Marie-Anne and many others). Also, a big Thank you! to the three musketeers (David Telford, Ricardo Pong-Wong, Alex Lam and Alex Clop – ok they are four) for their help, advice, discussions, proof reading but mainly for their friendship.

Thanks to all my friends and flatmates, that literally help me to survive in the UK, cooking, laughing and making the time very enjoy full (Simone, Ana, Lucio, Fede-Ale, Nikos and Chrysa); Without forgetting a thank you to my far away but always there friends (Aleyda, Fabby, Marchy, Melly and Irving). Also I want say a big Gracias! to Fede and Toby, that unfortunately for them, the rest left to sunnier places; therefore they carried all the heavy weight at the end, thanks guys!

Finally, I want to express my enormous gratitude and dedicate this to the two persons that I admire and love the most in my life, to my parents, Ma. Elizabeth Cardenas Cerda and Fco. Javier Cabrera McGregor, without them this would never be possible. Muchísimas gracias, porque sin ustedes y sin su apoyo y amor incondicional nunca hubiera realizado esto.

List of Publications

Refereed:

de Koning, D. J., C. P. Cabrera, and C. S. Haley, 2007 Genetical Genomics: Combining Gene Expression with Marker Genotypes in Poultry. *Poultry Science* **86**: 1501-1509

Conference Abstracts:

CP Cabrera, CS Haley and DJ de Koning 2007 Integrating technologies for the Analysis of Complex Traits. *Otto Warburg - International Summer School and Workshop 2007 on Computational Systems Biology.*

CP Cabrera, CS Haley and DJ de Koning 2007 Integrative and comparative genomics to target potential pathways for a functional body weight QTL in poultry. *Pathways, Networks, and Systems, International Conference.*

CP Cabrera, I Dunn, M Fell, P Wilson, DW Burt, D Waddington, R Talbot, PM Hocking, A Lam, A Law, CS Haley, S Knott & DJ de Koning 2007 Genetical Genomics for a marked QTL. *International Chick Meeting 2007*

CP Cabrera, I Dunn, M Fell, P Wilson, DW Burt, D Waddington, R Talbot, PM Hocking, A Law, CS Haley, S Knott & DJ de Koning 2006 Application of genetical genomics to a marked QTL in poultry. *8th World Congress on Genetics Applied to Livestock Production.*

Submitted (corrections phase):

CP Cabrera, I.C. Dunn, M. Fell, P. Wilson, D.W. Burt, D. Waddington, R. Talbot, P.M. Hocking, A. Law, S. Knott, C.S. Haley & D.J. de Koning. Complex Traits Analysis of chicken growth using Targeted Genetical Genomics

In preparation:

CP Cabrera, CS Haley and DJ de Koning. Exploiting microarray results using Integrative and Comparative genomics in livestock

List of Figures

Chapter 3

Chapter 4

Chapter 5

List of Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Abbreviations

BLAST	Basic Local Alignment Search Tool
bp	base pair
cM	CentiMorgan
db	Database
DDBJ	DNA DataBank of Japan
EMBL	European Molecular Biology Laboratory
*e*QTL	Expression quantitative trait loci
FDR	False discovery rate
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
ID	Identifier
IPI	International Protein Index
JIPID	Japan International Protein Information Database
kbp	kilobase pair (1000 bp of DNA)
LANL	Los Alamos National Laboratory
Mbp	megabase pair (10^6 bp of DNA)
MEA	Modular Enrichment Analysis
MIPS	Munich Centre for Protein Sequences
NBRF	National Biomedical Research Foundation
NCBI	National Center for Biotechnology Information
OBO	Open Biomedical Ontologies
PIR	Protein Information Resource
QTL	Quantitative trait loci
RH	Radiation Hybrid
SEA	Singular Enrichment Analysis
SGD	Saccharomyces Genome Database

Chapter 1 General Introduction

1.1 Bioinformatics

Bioinformatics is defined as the interdisciplinary field involving biology, computer science, mathematics, statistics, and biochemistry to analyze biological data, genome content, and to predict the function and structure of genes and macromolecules. Bioinformatics has developed from dealing with storage and data administration to specialized branches or areas of study. These include sequence analysis, proteome analysis, development of algorithms and computational tools, creation of advanced databases, text mining techniques, network prediction, interaction of molecules and pathway analysis. Bioinformatics has now extended its application to diverse research areas, ranging from medical applications (e.g. molecular medicine, drug development and antibiotic resistance) to even meteorological (e.g. climate change studies) and agricultural purposes (e.g. crop improvement and insect resistance).

In recent years, a major field for bioinformatics has been the investigation of the genome. The availability of genome sequences facilitates the study of genetic variability between and within organisms. Although functional gene characterization and prediction can be possible through exploiting bioinformatics methods and databases in order to identify the paralogs, protein families, and orthologous genes, it stills represents a major challenge (Andersen *et al.* 2008; Sonmez *et al.* 2009).

1.2 Complex traits and Genetical Genomics

One of the main objectives in genome research is to map and characterize trait loci that control variation in various phenotypic characters (e.g. to characterize genes that control growth, energy metabolism, development, appetite, reproduction and behaviour). These traits are commonly known as complex traits, and are considered to have a multifactorial background controlled by an unknown number of quantitative trait loci (QTL) as well as many environmental factors (Andersson 2001). Compared to traits which are controlled by a single gene (monogenic or Mendelian traits), complex traits usually reflect many small phenotypic contributions of multiple genes.

The phenotype modifications can occur due to various molecular events, such as single-nucleotide polymorphisms, multiple nucleotide variants in single genes that influence protein levels, small and large sequence deletions, and also polymorphisms on the coding regions or in the regulatory non-coding regions (Glazier *et al.* 2002). No gene acts on its own; each gene (and/or their protein products) interacts with many other genes, proteins and pathways, complicating the dissection of the molecular basis, even for the monogenic traits. The interaction of the causative gene with other products can reflect variation on the effects of the same trait, for example they can cause variable symptoms of patients with the same disease. Online Mendelian Inheritance in Man (OMIM) statistics as of March 2009 registered approximately 2,492 phenotypes with molecular basis known, and ~1,724 Mendelian phenotypes which remain with an unknown molecular basis. Another ~2,050 phenotypes are expected to be monogenic, though, it is hard to calculate the proportion of monogenic traits with known molecular basis which studied the 'chain reaction' and the effects of the interacting genes. It is believed that many of the 'declared' monogenic diseases will turn out to be

complex, possibly due to the unpredictable effects of gene mutations on the encoded proteins and the pathways where the proteins act (Peltonen & McKusick 2001).

Although QTL mapping is widely used to detect genetic regions responsible for phenotypic traits, the identification of the functional mutation and molecular basis of complex traits has only been successful for a very small proportion of QTL (Ron & Weller 2007). A remarkable example is a study made in pigs, where a polymorphism on an intron of *insulin growth factor 2* (IGF2) causing a major QTL effect through a subtle control of gene expression was identified by using genetic analysis (Van Laere *et al.* 2003). Another interesting example was presented by Clop *et al.,* (2006). This research localized a G-to-A mutation in the 3'UTR of the *myostatin* gene (GDF8) contributing to muscular hypertrophy explaining between ~ 20 - 33 % of the difference between parental sheep breeds.

A promising approach to obtain a better understanding of the genetic mechanisms influencing complex traits is genetical genomics, which brings together traditional QTL mapping with gene expression studies. In genetical genomics, the expression level of each transcript is treated as a quantitative phenotype and the marker genotypes are used to map loci affecting the gene expression levels, known as expression QTL (*e*QTL)(Jansen & Nap 2001). The integration of these technologies assumes that the gene expression levels are also affected by the functional polymorphism that affects the trait of interest (Arbilly 2006). The idea is to use segregation and recombination of related individuals where each individual of the population is used for genetic mapping and gene expression analysis. The gene expression profiling of all the individuals in a segregating population allows the expression level of each transcript to be treated as a quantitative trait for QTL mapping. The

*e*QTLs can be described as '*cis*-acting' when the *e*QTL is located in the same region as the gene that is affected, and as '*trans*-acting' when the *e*QTL and affected gene are not closely linked (Jansen 2003).

Cis-acting *e*QTLs are empirically found to be more significant. This is possibly because *cis*-*e*QTLs have larger effects on transcription (Gibson & Weir 2005). Additionally, *cis*-acting genes underlying a functional QTL by definition can be considered as key positional candidates for the functional QTL (Liu *et al.* 2001). One of the first approaches to integrate QTL studies with gene expression profiling was presented by Liu *et al.*, (2001). Basically, the use of comparative genomics facilitated the identification of causative genes of a QTL underlying resistance against Marek's disease in chicken. The goal was to identify positional candidate genes under the region of the QTL (*cis*-acting *e*QTL) that could be identified in the human genome. QTLs were mapped in an F_2 cross, while an expression study was carried out between the founder lines. Differentially expressed genes in the founder lines that co-located with QTL were positional candidates for these QTL. Fifteen of these genes were mapped onto the chicken genome, and twelve had an orthologous gene on the human genome. This methodology allowed the identification of two positional candidate genes for the QTL. One of the identified genes was the growth hormone (GH1). More interestingly, there is supporting evidence based on literature for this positional candidate gene to be biologically involved with Marek's disease resistance (Liu *et al.* 2001).

However, some other recent studies found only a small proportion, or complete lack, of *trans*-acting *e*QTL effects, suggesting that most of the *e*QTL are in or near the gene whose variation they explain (Pastinen *et al.* 2006). Usually *trans*-acting *e*QTLs explain less than 20% of the phenotypic variance; also, they show smaller effects than *cis*-acting *e*QTls, resulting in lower

statistical power and typically falling under the threshold of detection for linkage studies (Petretto *et al.* 2006).

Another behaviour of *e*QTL studies that has been observed is the clusters of *trans*-acting *e*QTLs affecting the expression of a much larger number of transcripts than expected by chance, commonly called 'hotspots' (Yvert *et al.* 2003). The basis underlying the detected hotspots, i.e. whether there is a true correlation reflecting the effect of gene regulation or a spurious correlation because of technical and/or environmental factors, remains unclear (de Koning & Haley 2005).

A drawback in genetical genomics is that the number of traits to be analyzed is very large, and because of their cost, the experimental sizes are relatively modest, resulting in limited power of published studies (de Koning & Haley 2005). Therefore, optimizing the statistical power of genetical genomics experimental designs is crucial. Alternative experimental designs based on comparative genomics, selective phenotyping, or optimal distant pairings have been investigated to improve the power of the gene expression experiments, under different hypotheses (Liu *et al.* 2001; Borevitz & Chory 2004; Fu & Jansen 2006; Rosa *et al.* 2006).

Despite the fact that genetical genomics is a relatively recent approach and the methods are still under development, it has been applied to different organisms like yeast (Brem *et al.* 2002; Yvert *et al.* 2003), eucalyptus (Kirst *et al.* 2004), maize (Schadt *et al.* 2003), mouse (Schadt *et al.* 2003; Bystrykh *et al.* 2005; Chesler *et al.* 2005), rat (Hubner *et al.* 2005), pig (Ponsuksili *et al.* 2005), chicken (Liu *et al.* 2001) and human (Schadt *et al.* 2003; Morley *et al.* 2004; Monks *et al.* 2004). Most of these studies were driven by the interest of chasing regulatory genes which might control complex networks and were

able to identify causal genes responsible for a certain trait or disease (Secko 2005). Schadt and colleagues, (2003) studied expression profiles obtained from 111 liver tissues from an F_2 cross population between two mice inbred strains (C57BL/6J × DBA/2J) in order to investigate genes linked to obesity. In their study they indentified 2,123 genes with differentially expressed profiles and 4 *cis*-acting genes linked to obesity. A similar example was presented by Hubner *et al*, (2005). Their interest was to study the regulation of gene expression in a recombinant inbred rat strains (BXH × HXB) to investigate hypertension. By using comparative mapping they were able to identify 73 candidate genes. Furthermore, by only identifying candidate regulatory genes, Bystrykh *et al*, (2005) and Kirst *et al*, (2004) made an effort to investigate pathways and genetic networks through the application of genetical genomics. The eucalyptus study assayed 2,608 genes of a *E. grandis* × *E. globules* backcross population to reveal the genetic networks responsible for growth variation. Two loci were discovered to coordinately control lignin biosynthesis. In addition, these two loci localized in growth related QTLs. The authors suggested that the targeted regions might regulate growth, lignin content and composition (Kirst *et al.* 2004). Some of these genetical genomics initial findings made the field very popular.

A great expectation from genetical genomics is the potential to successfully reconstruct gene networks by developing advanced algorithms and making possible the integration of data from multiple sources (e.g. genotypic, molecular and expression profiling). These integrative approaches could facilitate the understanding of the underlying causes of complex traits and diseases (Schadt *et al.* 2005). Another promising expectation is that genetical genomics could lead to a better understanding of behavioural and stress response traits in animals. These often include non genetic factors such as

feed availability, comfort, temperature, and humidity. Expression data on animals exposed to different environmental conditions could show how animals adapt to the environment (Kadarmideen *et al.* 2006).

1.3 Bioinformatics for the dissection of complex traits

The bioinformatics tools and annotation procedures attempt to simplify the methods and analyses used for the genetic dissection of complex traits. The integration and organization of gene expression data, functional genomics, pathways and molecular biology among others, has driven genetics into a integrative genomics approach. This approach attempts to achieve a holistic approach, instead of exploiting each field independently. By combining the different sources of information it is possible to obtain better insights, which would not be achievable by individual interpretations. The integrative genomics approach helps scientists to reduce significantly the list of candidate genes, and identify the cellular and metabolic pathways and the downstream effects that contribute to complex traits and diseases.

Data integration faces important challenges, where the capability of analyzing large datasets and bringing them together has not developed equally. One issue is that the data was created from different domains and each domain has its own ways of access and storage. The information identifiers (accessions), terminologies and formats vary from source to source, complicating the process of querying and accessing the data. Regardless of all the attempts to unify the information, these continue as an unsolved puzzle. When analysing large-scale studies, the analysis process becomes even more complicated, as the differences will represent major difficulties on the automating and standardization processes. Another aspect

to be considered is that the bioinformatics tools available for genome research are highly variable from species to species. For example, in non-model species the analyses are very complex because of the lack of annotations, direct experimental data and pathway associations. In these situations, the use of extensive comparative genomic approaches can exploit the information that has been gathered on model organisms.

In recent years, some emphasis has been placed on investigating methods and advanced algorithms for the development and integration of systematic strategies to exploit 'all' data in comprehensible ways. Several studies focus on functional discovery rather than gene discovery. Fisher *et al.,* (2007) presented a systematic strategy to investigate genotype-phenotype correlations to identify candidate genes underlying complex traits. The strategy focuses on the analysis of the QTL and gene expression data at pathway levels, in order to emphasize on functional discovery. Firstly, the genes under the QTL region are defined through the limits of the physical location of the QTL. Then, these genes are annotated and linked to their 'known' pathways using KEGG pathway database. The differentially expressed genes obtained from the microarray study follow the same process as the genes under the QTL region, creating two sets of lists of pathways. Finally, a subset of common identified pathways is obtained, allowing the detection of those processes that might be influencing the phenotypes (Fisher *et al.* 2007). Compared to 'traditional' gene hunting and gene-level analyses, the functional discovery approaches will provide a broader view of the process involved in complex traits.

Thesis motivation

There is currently a great interest in the investigation of systematic methods for the dissection of complex traits. However, the exploration and integration of the available data originating from the various experimental areas in an accurate and automated way has not been achieved. In order to exploit the data and make it more interpretable and useful for science, we need a systematic way to integrate and analyse to the maximum potential the results generated by quantitative trait analyses, microarray studies and molecular biology.

Thesis objective

The objective of the thesis is to investigate bioinformatics methods, tools and frameworks in order to obtain a better understanding of the mechanisms governing complex traits. The aim is to be able to characterize identified QTL in the best possible way. For example: what genes are likely to be responsible for a certain trait; how these genes interact with each other; ideally, what is the proportion in which the environment affects the phenotypic traits; what are the downstream (global) effects; what other QTLs have been found for the same trait in the same species and also in other species; what QTLs have been reported in the same region where the identified QTL was reported.

The thesis attempts to get closer to the biological interpretation of high-throughput data and the genetic characterization of QTLs by exploiting various sources of information and bringing them together, and ultimately to target potential candidate genes that could be responsible for a certain phenotypic trait.

Thesis Overview

Shown below is an overview of each chapter:

Chapter 2 Introduce the concepts of microarrays, including normalization and statistical methods, technical issues, and the challenges microarrays and their analysis face. This chapter also presents a review of the current status of bioinformatics tools and databases available for genome research, discussing briefly the procedures behind the tools.

Chapter 3 The concept of genetical genomics and their experimental designs is extended. A pilot study made in chickens (between broilers and layers) based on a targeted genetical genomics approach is presented and analysed. The targeted approach contrasts the gene expression levels of two alternative genotypes of a known QTL, hence narrowing the QTL-phenotype gap. The study allows identification of positional candidate genes and genes with high expression variability between the two genotypes. Additionally, the manual annotation procedure is also presented.

Chapter 4 One of the biggest limiting factors in the investigation of any high-throughput technology is the annotation. In model organisms, the annotations are more complete than those in livestock species. The methods of how the annotation procedures can be improved and utilized for the genetic dissection of complex traits is investigated. Subsequently, the annotation framework used for the analysis of the body weight study of the targeted genetical genomics approach is discussed.

Chapter 5 Post-analyses results and methods are exemplified through the targeted study. Also, an example of how the use of comparative genomics and the integration of technologies can drive us to a better characterization of an identified QTL.

Chapter 6 Features the concluding remarks of the thesis. A final critical opinion on the present and future bioinformatics methods used for the analyses of complex traits is discussed.

Chapter 2 Bioinformatics Procedures and Tools

2.1 Introduction

The rapid development of high-throughput technologies and their constant decreasing costs results in the production and availability of massive amounts of data of various types. One of the most commonly used and well studied technologies is gene-expression microarrays. Microarrays have spawned debates in many areas, from their technical quality measures to the application of appropriate statistical methods. Microarrays are further discussed in the first section of this chapter.

As the amount of genetic data increases very rapidly, the bioinformatics procedures to analyse these data has not progressed at the same pace. Although many tools have been developed and the number of tools continues to increase, analysing the data and choosing the appropriated programs for the analyses still represents a considerable challenge. Additionally, many available resources for different areas (i.e. network modelling and text-mining) are still under development and one should understand the algorithms of the programs and the kind of results that they can provide. The amount of tools available and their functional similarity can confuse a researcher. It becomes difficult to remember and understand the differences between the applied algorithms and the way each application works (i.e. steps used to utilize the programs). A brief description of bioinformatics procedures and methods used in this thesis is presented in this chapter, together with some available tools and genomic resources.

2.2 Microarrays

An initial promise of microarray studies was to narrow the gap in understanding gene functions and molecular mechanisms (Brown & Botstein; Lockhart & Winzeler 2000). Since its initial appearance, microarray technologies have constantly improved and developed their applications and techniques. Although the most common application of microarrays is the gene expression analysis, recently this technology has been applied to detect single nucleotide polymorphisms (SNPs), alternative splicing, miRNAs and in the investigation of evolutionary and epigenetic studies (Hoheisel 2006; Yin *et al.* 2008).

Microarray technology allows the monitoring of gene expression levels on a large scale. Thousands of features (e.g. cDNA, oligonucleotides) are attached onto a solid surface (e.g. glass or silicon slides) at fixed locations (spots), each of them representing a gene transcript. The targets are labelled with a fluorescent dye(s) for hybridization. To measure the RNA abundance, the arrays are excited with a laser light, where the fluorescence intensities estimate the relative expression levels of the transcripts (Brazma & Vilo 2000).

The goal of microarray studies can be classified as class comparison, class discovery and class prediction. The 'classical' design is the class comparison, where the goal is identifying differentially expressed genes between two or more groups (e.g. control vs. disease). Class discovery refers to the identification of patterns or groups within the samples (arrays) and/or gene expression levels. Finally, class prediction involves the prediction of group membership for a given sample based on the gene expression profiles (Olson 2006).

2.2.1 Microarray Platforms

There are several types of arrays available, such as cDNA arrays, long-oligonucleotide arrays (e.g. Agilent) and Affymetrix Gene-chips (Ness 2007). The chosen platform for the microarrays can have a great impact on the analysis. Two colour platforms co-hybridize selected pairs of samples. An advantage of one colour platforms is that they give a direct measurement for a single sample and can be extended very easily (Fu & Jansen 2006). Affymetrix is the most popular commercial single channel microarray platform. These arrays consist of short oligos (~24 – 85bp), and are highly standardized facilitating an easier comparison even between experiments. Another advantage of single channel arrays is that the probes on the arrays are created to represent unique genes (Shiu & Borevitz 2006). In two colour spotted cDNA microarrays, the sequences are approximately larger than 300 bp making hybridization more reliable and likely to identify transcripts with alternative splicing events, but without distinguishing between splice variants. Usually, two mRNA samples are transcribed into cDNAs, labelled with different fluorescent dyes (most commonly with *Cy*3 and *Cy*5) and then hybridised onto the same slide. The dyes are measured separately and captured into independent images. These arrays allow more flexible designs, and in principle they can provide double the amount of data provided by one colour platforms. The major advantage of these arrays is their manufacturing cost is less expensive than other methods and also that they are easier to prepare. The advantages of these arrays also represent some of the disadvantages. The larger sequences on the arrays represent a higher risk of cross-hybridization of the probes with higher identity. Further, a low quality sample might highly influence the measurement of the other sample. In addition, the spotted arrays are more prone to quality problems, where

the shapes, intensities and variations among the features are not uniform (Shiu & Borevitz 2006). In order to remove some of the systematic variation, the images are scanned and normalized before the expression values are analysed (Yao *et al.* 2004; Allison *et al.* 2006). Image analysis and data extraction are technically explained in more detail by Duggan *et al.,* (1999).

2.2.2 Experimental Designs

The processes of experimental design, data accessibility, and platform selection usually precede the analysis of the microarray and they have a direct impact on the statistical analysis of the data. Important aspects to take into consideration in the experimental designs are the type of experiment and the use of replicates. The experimental design is strongly linked with the goal of the study, a good design should aim to reduce variation sources and gain the most information possible with the minimum use (number) of arrays, therefore also reducing the experiments costs (Leung & Cavalieri 2003).

The use of biological and technical replicates has been considered as an essential step, which could increase the power of the experiment and can help to increase the proportion of true differentially expressed genes among significant results. A technical replicate refers to the same sample being hybridized to different arrays; and biological replicates are those where different individuals have been selected for samples (Brazma & Vilo 2000; Allison *et al.* 2006; Olson 2006).

The idea of pooling biological replicates has been considered to reduce variability among arrays and also reduce the overall cost of the experiment. This is only relevant where there is no interest in the individual replicates,

but only in identifying differentially expressed genes between different groups (Allison *et al.* 2006).

The way the samples are paired on a two colour array is assigned according to the experimental design selected. Some of the most common microarray experimental designs are the reference and loop designs. The reference design utilizes one channel (one dye) of each array as a 'control reference' using a reference pool of RNA, and the other channel is to hybridize the samples of interest, which assumes there are no dye-effects affecting the targets on the array. The reference designs require n number of arrays to assess n number of samples. The loop design was proposed to minimise the number of arrays required for n samples. In the loop design, each target is labelled with both dyes (*Cy*3 and *Cy*5) on different arrays forming a *loop*. For example, sample x is hybridized onto array 1 with *Cy*3 paired to sample y - *Cy*5; array 2 would pair sample y- *Cy*3 with sample z-*Cy*5; array 3 would pair z-*Cy*3 paired to x-*Cy*5 (i.e. array 1 (x,y), array 2 (y,z), array 3 (z,x)). Therefore the array dye-effect would be accounted for, although assuming that there are no gene-specific dye-effects. An advantage of the *loop* design is that it would require half of the arrays as the samples being measured (Kerr & Churchill 2001). A dye-swap experiment is an experimental design which can minimize systematic bias and 'ensure' correction of gene-specific dye-effects. The hybridization is done twice for each sample, exchanging dyes on the second hybridization (Yang & Speed 2002). Following the previous notations, x-*Cy*3 hybridized to array 1 paired to y-*Cy*5; and array 2 hybridized with y-*Cy*3 together with x-*Cy*5 (i.e. array 1 (x,y) array 2 (y,x)). More specialized experimental designs are presented in Chapter 3.

2.2.3 Normalizations

Normalization of the expression values by adjusting the spot intensities is an essential step in order to make the arrays comparable. This process attempts to control systematic variation among the experiment(s) (Quackenbush 2001; Butte 2002). Normalization can be performed in several ways but the most 'appropriate' model for normalization is still under debate (Allison *et al.* 2006) and could well vary from slide to slide.

Yang *et al.*,(2006) distinguished between three approaches for normalization and described the methods that could be applied in those cases: 1) The first case 'within-slide' adjustments are undertaken independently for each slide, where global normalization, intensity dependent normalization and within-print-tip-group normalization are some of the approaches that can be applied. Global normalization assumes a constant factor relating the intensities, centring the distribution of log-ratios to zero. Intensity dependent normalization uses lowess (loess): a statistical smoothing technique that performs locally linear fits where dye bias tends to be caused by spot intensity. Within-print-tip-group normalization attempts to reduce variation caused by print-tips on the spotting device; 2) 'Paired-slide' normalization applies in the case of designing an experiment with dye-swap, adjusting only print-tip (pins) locations, and this approach assumes similar log-ratios distribution on the two slides; 3) In the third case, 'multiple-slide' normalization attempts to assign a normal scale among arrays to make them comparable, regardless of the type of within-slide normalization process (Yang *et al.* 2006).

In practise, a global normalization method is frequently used despite the evidence of systematic regional bias within slides (Qiu *et al.* 2005; Reimers & Weinstein 2005).

Normalization models attempt to reduce the 'noise' caused by technical systematic errors. However, Qiu *et al.,*(2005) demonstrated that the normalization could not only remove artificial correlations but they can also affect the true correlation of gene interactions. Therefore, normalizations can have an impact on the detection of molecular pathway and gene regulation networks, where the clustering approaches depend on the correlation structure among gene expression levels.

2.2.4 'Significant' Gene Identification

In early studies the fold change cut-off was used as a measurement to identify differentially expressed genes. However, it does not produce known and controllable error rates and therefore the use of the fold change on its own is not reliable (Allison *et al.* 2006). The most commonly used methods to infer that the genes are differentially expressed are: 1) *t*-test, to determine statistically 'significant' difference between two groups by looking at the difference between two independent means; 2) Analysis of variance (ANOVA), a general statistical technique to identify significant differences between two or more groups; 3) logistic regression, a technique used when the outcome variable is binary; and 4) survival analysis, used to analyse time to event data (Allison *et al.* 2006).

In the case of selecting a comparison class type of experiment for comparison of two conditions, a two-group test such as a *t*-test is suitable. In the case of handling an experiment with multiple conditions an analysis of variance (ANOVA) would be more appropriate. The type of ANOVA will depend directly on the experimental design (Olson 2006).

With the idea of testing thousands of genes simultaneously, biologists have to allow that some proportion of the significant results will be false.

Benjamini and Hochberg (1995) introduced the concept of false discovery rate (FDR). Broadly, the FDR represents the proportion of false positives (referred to as the *q-value*) among those results that are called significant given a certain *p-value* (the probability of a false positive for a single test). The use of FDR approaches means 'accepting' some false positives while it makes an effort to control the extent of them (Reiner *et al.* 2003; Grant *et al.* 2005). Biologists have adopted the use of the FDR in their studies as an alternative to control false positives using the much more conservative Bonferroni correction. Storey and Tibshirani (2003) exemplify with several genome-wide studies how the FDR measure manages an acceptable balance between the number of true and false positives.

2.2.5 Validation

Microarray experiments are highly susceptible to accumulated errors during an experiment ranging from batch effects to bias introduced by insufficient normalization. Consequently, extreme care should be taken in each step, from the experimental design to the biological annotation of the probes in the microarray and conclusions drawn from the study. Validation is the process of proving the veracity of the study. Generally, validation can be classified in operational and constructive. Operational validation is when the hypotheses should be re-tested using the original methodology, when the hypotheses are tested by different means, this is known as constructive validation or constructive replication (Allison *et al.* 2006).

2.2.6 Challenges

While microarray manufacturing and production techniques are well defined, many studies have focussed on the processing and analysis steps. Despite vast numbers of publications, many questions remain: A) what type

of experimental design is most informative and cost efficient? B) How to assess the quality of the experiment? C) What type of normalization is most convenient? D) What statistical models and thresholds are the right ones to consider whether genes are differentially expressed? E) How can regulatory networks be inferred? and F) How could clustering methods be applied to expression levels?

2.3 Bioinformatics Methods and Procedures

This section presents in brief the theory behind the bioinformatics procedures that have become essential and part of more complicated analyses in the investigation of high-throughput technologies. The procedures discussed in this section are sequence and enrichment analyses. The use of these methods aims to characterize and investigate functional and evolutionary mechanisms of the originated genetic and biological data.

2.3.1 Sequence Analyses

Dayhoff and colleagues pioneered the investigation on computational sequence analyses during the 1960's. They organized and studied protein families based on sequence similarity, giving rise to the idea that similar sequence proteins might conserve similar biochemical functions and three-dimensional structures. Additionally, to investigate further sequence substitutions (dissimilarities between sequences) they elaborated a set of tables (matrices) containing the probabilities of amino acid substitutions (e.g. PAM) and graphically display the sequences according to 'similarity scores' on structured 'trees' (now commonly known as phylogenetic trees) (Mount 2001) .

2.3.2 Sequence Alignments

Sequence alignments are useful for functional, structural and evolutionary discoveries. A sequence alignment is defined as the procedure of comparing two (pair-wise alignment) or more (multiple sequence alignment) sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences.

Comparison, between sequences can be performed in two ways. One method involves finding the optimal alignment across the entire length of the sequences, known as global alignment. The other method is known as local alignment, which focuses in the identification of sections of the sequences with very strong similarity. The global alignment can be useful in cases where the sequences are similar between each other and also are approximately the same length. The local alignment is useful when looking for conserved regions or very similar patterns.

An important observation and a very common mistake is the confusion of some terms typically used in sequence analyses, such as the difference between sequence similarity and homologs. Sequence similarity, or sequence identity, is simply the score of the matching characters in an alignment, whilst homologs refer to a common evolutionary origin (e.g. genes descended from a common ancestor). Homologs genes that are related through gene duplication events are known as paralogs, but if the genes are derived through speciation (or vertical descent) they are referred to as orthologs.

2.3.3 Enrichment Analyses

Enrichment analyses contribute to functional characterization and analyses of large gene lists. These analyses categorise the genes into similar functions.

Pathway and gene ontology analyses comprise the research of variation expression patterns in previously defined classes of genes (e.g. cytoskeleton, apoptosis, membrane transport, metabolisms). Usually, these methods are based in either text mining or scientific literature research, helping the users to identify and discover novel and unnoticed interactions between differentially expressed genes (Cavalieri & De Filippo 2005; Olson 2006). Although the enrichment tools utilize diverse methods, the general structure is composed by three major features, the databases or backend annotations, the algorithms and statistical procedures, and the visualization and exploration of results (Huang *et al.* 2009).

There has been a large increase in the availability of tools in this area in recent years. In 2005, Khatri and Draghici (2005) were able to collect and review 14 enrichment tools, by 2009, Huang and colleagues reviewed 68 applications. They classified the tools according to the algorithm utilized and categorized them in three classes (singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA)). SEA is the 'traditional' strategy of iteratively testing enrichment of each annotation term of 'significant' gene lists. GSEA evaluates the entire gene list (no cut-off threshold required) and expression values which are integrated into the *p*-value calculation (Subramanian *et al.* 2005). MEA is also based on the SEA incentive, although a powerful feature is that it includes term-term and/or gene-gene relationships into the calculation of the *p*-value (Huang *et al.* 2009).

The goal of the enrichment analysis is to indentify statistically significant functional categories. This is performed by estimating if x gene shows overrepresentation in a certain category. The calculation results are complicated because of diverse factors, such as, gene lists size, gene

annotation overlaps, strong relationships among genes, and unbalanced distribution of the annotation (Huang *et al.* 2009). Goeman and Buhlmann discuss in more detail the methodological statistical issues and assumptions researchers face when analysing gene set enrichments. Typically, the statistical algorithms behind these applications apply four different methods (Chi-square χ^2, Fisher's exact test, binomial probability, and hypergeometric distribution) (Khatri & Draghici 2005). The hypergeometric distribution is applied to calculate the probability of a functional category occurring *x* times by chance in a 'significant' gene list. However, this method is not easily applicable for arrays containing a large number of features. In such cases, the binomial model is the method that tends to be used (Tavazoie *et al.* 1999; Draghici *et al.* 2003). The χ^2 and Fisher's exact test are used for equality proportions. χ^2 describes how the observed number of features (genes) deviates from what is expected. Although χ^2 should not be applied in cases where the features are less than five per category (significant genes on a same pathway), the Fisher's exact test could be an alternative solution (Draghici *et al.* 2003).

Furthermore, some tools utilize what is known as the 'hit-counting' method (i.e Z scores and odd-ratios). The Z score (standardized difference score) is based on the hypergeometric distribution. The calculation of the Z score takes into consideration four factors: 1) the total number of genes on the microarray; 2) number of genes per pathway; 3) total number of genes differentially expressed at a certain threshold, and 4) the number of significant genes found per pathway (number of hits). The results are interpreted as when the score is zero there was no enrichment found, a positive score represents enrichment and a negative value under representation. It has been argued that the Z score is diffuse. The odds-ratio

approach calculates the probability of a GO term appearing on the gene list divided by the chances of appearing on the entire GO category (higher level GO category) (Curtis *et al.* 2005).

In addition to the *p*-value calculations and enrichment scores, some tools include correction of multiple experiments. Such a correction should be used when there is no a *priori* functional category and many categories are evaluated simultaneously. Although the Bonferroni and Šidák adjustments make a strong and false assumption (that the variables are independent), these methods are suitable when not many functional categories are assessed. The false discovery rate (FDR) would be more appropriate when the functional categories are known to be related, for very strong relationships with enough categories to perform simulations the Monte Carlo or bootstrap approach would be more useful (Khatri & Draghici 2005).

2.4 Bioinformatics Tools

Publicly available resources provide different types of information about genes and their gene products. The following section presents a brief description of 35 tools and bioinformatics resources evaluated/used during the thesis.

2.4.1 Sequence Analysis Tools

Basic Local Alignment Search tool (BLAST; http://blast.ncbi.nlm.nih.gov/Blast.cgi).

This tool was developed to find (local) similarity between two sequences. Given a sequence of interest this program searches against protein and

nucleotide sequences databases and calculates the statistical significance of the alignments results (Altschul *et al.* 1990). The BLAST algorithm is hosted by the NCBI, and contains a large number of organism sequence databases (approximately 103 organisms as of February 2009). The BLAST family is formed by BLASTN, BLASTP, BLASTX, TBLASTN, megaBLAST and psi-BLAST. These programs allow use of input nucleotide, translated or protein sequences as queries and searches against various databases. A very similar algorithm also widely used is the BLAT algorithm. This algorithm is considerably faster and presumably more accurate than BLAST (Kent 2002).

CLUSTALW (http://www.ebi.ac.uk/tools/clustalw2)

CLUSTALW is one of the most commonly used programs to perform multiple sequence alignments. This program allows the performance of global multiple nucleotide and protein weighted sequence alignments, and was recently re-programmed to provide faster and more precise results (Larkin *et al.* 2007).

Torniainen *et al.,* (2009) utilized this method to study the conservation sites across species of the LCT gene involved in congenital lactase deficiency, and successfully targeted four novel mutations.

Geneious (http://www.geneious.com/)

Geneious software (Copyright © 2005-2009 Biomatters Ltd.) allows performing integrated DNA and protein sequence analysis, BLAST and access to public databases. One of the most advantageous features of the software is the sequence alignments manageability (both pair-wise and multiple sequence alignments) and visualization; also the facility of sequence translations and open reading frame (ORF) findings. Additionally, the

alignments produced can be computed to view them graphically as phylogenetic trees.

2.4.2 Databases and Genome Resources

The first database to appear was a protein sequence database (Protein Identification Resource (PIR)) in the 1960's. Following this, the Munich Centre for Protein Sequences (MIPS) and Japan International Protein Information Database (JIPID) were created and worked together to become the PIR-International database maintained by the National Biomedical Research Foundation (NBRF; http://pir.georgetown.edu/nbrf/). This was formed in 1984, and recently merged into the UniProt database (http://www.uniprot.org/).

The first DNA sequence databases were the GenBank database, under the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) and the European Molecular Biology Laboratory (EMBL; http://www.embl.org/). Afterwards the DNA DataBank of Japan (DDBJ; http://www.ddbj.nig.ac.jp/) was created and together with the NCBI and EMBL formed the International Nucleotide Sequence Database Collaboration (http://www.insdc.org/).

A crucial change which revolutionized the way research and biological information was shared, managed and investigated was the availability of the databases as freely available worldwide resources. This allowed querying the main databases through single programs and interfaces based on the internet. In 1993, the *Nucleic Acids Research* journal start publishing an annual issue dedicated to the available databases. The first accessible issue from the internet was published in (2004). The current publication (2009) holds information on 1170 databases (Galperin & Cochrane 2009). Each year for the

last 6 years, approximately 130 new databases were included in the system (Table 2.1). The records have been divided into 14 main database categories. Appendix 2.1 presents a list of the main categories and their sub-categories with the total number of databases found per class. A full list of all the databases, their summaries and updates can be found online at the *Nucleic Acids Research* web site (http://nar.oxfordjournals.org/).

Table 2.1 Database records on Nucleic Acids Research Database issue

Publication year	Total Number of Databases	Added	Withdrawn
2004	548	162	
2005	719	171	17
2006	858	139	3
2007	968	110	11
2008	1078	110	25
2009	1170	92	16

The following section briefly describes the NCBI, EBI-EMBL and UCSC features and systems. These are considered to be some of the major data and tool providers for the biological, molecular and bioinformatics research communities. Additionally, the RGD database and AnimalQTLdb are also described.

NCBI (http://www.ncbi.nlm.nih.gov/)

The NCBI centralises various specialised databases (divided into molecular and literature), and offers a wide range of bioinformatics tools. Within the molecular databases we can find nucleotide, protein, structure, taxonomy, genome, expression and chemical databases. The literature database comprises research articles (e.g. PubMed), and collections of reference overviews of Mendelian disorders (OMIM, OMIA). BLAST is one of the tools housed by the NCBI, and the genome map viewer and ORF finder are other

tools also available. Table 2.2 summarises the main NCBI databases and tools.

Table 2.2 NCBI resources

NCBI RESOURCES		
NAME	URL	**DESCRIPTION**
Genbank	http://www.ncbi.nlm.nih.gov/Genbank/	Genetic DNA sequence database. Holds approximately 82,853,685 sequence records as of February 2008.
UniGene	http://www.ncbi.nlm.nih.gov/unigene	Organizes sequences into a non-redundant set of gene-oriented clusters, where the gene name is set as the title of the clusters.
Entrez Gene	http://www.ncbi.nlm.nih.gov/Entrez/	Provides a single-query interface retrieval system to NCBI main databases.
MeSH	http://www.nlm.nih.gov/mesh/meshhome.html	Provides a consistent way to retrieve information that may use different terminology for the same concepts.
OMIM	http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim	Catalogue of human genes and genetic disorders.
OMIA	http://www.ncbi.nlm.nih.gov/sites/entrez?db=omia&tool=toolbar	Database of genes, inherited disorders, and traits in animal species.
HomoloGene	http://www.ncbi.nlm.nih.gov/homologene/	A gene homology tool that compares nucleotide sequences between organisms, providing automated detection of homologs among the annotated genes of several eukaryotic genomes.
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	Conserved Domain Database contains protein domains imported from Pfam and SMART
Map Viewer	http://www.ncbi.nlm.nih.gov/mapview/	Provides genome mapping and sequencing data of various organism

EBI-EMBL (http://www.ebi.ac.uk/)

The European Bioinformatics Institute (EBI) research centre and bioinformatics services form part of the EMBL, and provides and hosts literature, sequence, microarray, pathway, networks and ontology databases

and tools. UniProt, ArrayExpress, Ensembl, InterPro and Biomart are some of the most recognized EBI resources (Table 2.3)

Table 2.3 EBI-EMBL Resources

EBI-EMBL RESOURCES		
NAME	URL	**DESCRIPTION**
EMBL	http://www.ebi.ac.uk/embl/	Nucleotide sequence database (EMBL-Bank).
UniProt	http://www.uniprot.org/	Universal Protein Resource, protein sequence and function database.
InterPro	http://www.ebi.ac.uk/interpro/index.html	Protein families, domains, regions, repeats and sites database.
ArrayExpress	http://www.ebi.ac.uk/microarray-as/ae/	Public repository warehouse for transcriptomics data. Holds approximately 7,570 experiments (February 2009) .
BioMart	http://www.biomart.org/	Query oriented data management system through a web interface, provides data mining searches for complex datasets.
Ensembl	http://www.ensembl.org/index.html	Genome browser, provides functional gene annotations (Sanger Institute project collaboration.).

UCSC (http://genome.ucsc.edu/)

The University of California Santa Cruz (UCSC) genome browser contains reference sequences and draft assemblies for a large number of genomes; additionally it provides tools such as Genome Browser, BLAT, Gene Sorter and Genome Graphs (Table 2.4). A very useful feature of the UCSC is the versatility of the Genome Browser which is able to display simultaneously a large amount of features over the chromosomes (e.g. mapping and sequencing tracks, phenotype and disease associations, comparative genomics, mRNA, EST, expression, regulation, variation and repeats).

Table 2.4 UCSC Resources

UCSC RESOURCES		
NAME	**URL**	**DESCRIPTION**
BLAT	http://genome.ucsc.edu/cgi-bin/hgBlat?command=start	Sequence alignment tool.
Table Browser	http://genome.ucsc.edu/cgi-bin/hgTables	Text-based access to UCSC Genome Browser databases.
Genome Graphs	http://genome.ucsc.edu/cgi-bin/hgGenome	Possible to upload and display genome-wide datasets (i.e. genome-wide SNP association studies).
Gene Sorter	http://genome.ucsc.edu/cgi-bin/hgNear	Contains various gene-relationships (i.e. gene expression, gene homologies).

Animal QTLdb (http://www.animalgenome.org/QTLdb/)

The animal quantitative trait locus database holds publically available QTL data on livestock species (pig, cattle, chicken and sheep). Additionally, the QTL data is linked to other genomic resources (i.e. radiation hybrid (RH) maps, physical maps and human genome maps). Currently (as of January 2009) the PigQTLdb counts includes 1,831 QTLs (316 different traits); CattleQTLdb holds information about 1,123 QTLs (101 traits); ChickenQTLdb 657 contains QTLs (112 traits); and, SheepQTLdb is an on-going work which at present holds 51 QTLs (27 different traits).

2.4.3 Enrichment Tools

A large number of tools have been developed with the purpose of analysing and interpreting biologically the functional enrichment of high-throughput datasets. Although there are many, the tools can differ in important features, such as installation capabilities and supported platforms, required identifier system, statistical methods, visualizations and presentation of results and, very importantly, sources of annotation and available organisms. Khatri and Draghici (2005) performed a detailed analysis on 14 enrichment tools taking

into consideration the annotation sources and statistical models. During the same year (Curtis *et al.* 2005) presented another review analysing 15 tools although some of these overlap with the ones presented by Khatri and Draghici. For a recent review on some of the available enrichment tools please refer to Huang and colleagues. Table 2.5 lists some of the enrichment tools that were used during the thesis.

Table 2.5 Enrichment Tools

Pathway/Functional/Network Analysis Systems Tools		
NAME	**URL**	**DESCRIPTION**
AgBase (McCarthy *et al.* 2006)	http://www.agbase.msstate.edu/	Tools to perform GO functional analysis for agricultural species. Programs used GOProfiler, GORetriever, GOanna, GOSlimViewer.
Ariadne Pathway Studio	http://www.ariadnegenomics.com/products/pathway	Builds, visualizes, analyses and curate pathways; Import and analyse gene and protein lists; interpret microarray gene expression data. Supports GO, Molecular signal database (GSEA), HPRD, KEGG, GEO, BioPax and OBO. Functional relationships available for human, rat and mouse, and 10 plant organims. (Commercial)
BioPAX	http://www.biopax.org	Biological pathways exchange collaboration. Contains information for metabolic pathways, molecular interactions, signaling pathways, gene regulation and genetic interaction. Resources BioCyc, KEGG, Reactome, INOH, Biomodels, Nature/NCI Pathway Interaction database, Cancer Cell map, Pathway common.
DAVID (Dennis *et al.* 2003)	http://david.abcc.ncifcrf.gov/	Bioinformatics resource for functional annotation, gene functional classification, gene ID conversion, gene batch viewer, visualization and integrated discovery of genomic results
GeneMAPP (Dahlquist *et al.* 2002)	http://www.genmapp.org	Application designed to visualize gene expression and other genomic data on maps representing biological pathways and GOs. Useful tool to identify

		metabolic pathways.
GSEA	http://www.broad.mit.edu/gsea/index.jsp	Gene Set Enrichment Analysis Tool, software statistically tests significance for prior defined sets of genes between two states.
iHOP (Hoffmann & Valencia 2005)	http://www.ihop-net.org/UniPub/iHOP/	Information hyperlinked over proteins, gene and protein interactions based on literature relationships.
IPA	http://www.ingenuity.com/	Ingenuity Pathway Analysis (IPA). A web-based application that allows users to search scientific literature, build dynamic pathway models, and analyze experimental data. (Commercial)
KEGG (Ogata *et al.* 1999)	http://www.genome.jp/kegg/pathway.html	Kyoto Encyclopedia of Genes and Genomes provides information about both regulatory and metabolic pathways for genes.
Ontologizer (Bauer *et al.* 2008)	http://compbio.charite.de/index.php/ontologizer2.html	GO software for statistical analysis and visualization of high-throughput data
PubGene	http://www.pubgene.com	Up-to-date information on gene and protein relationships from the literature that can be used to put microarray results in the context of possible new interactions or gene networks. Core program freely available, advanced features of the program are commercial.
Reactome (Vastrik *et al.* 2007)	http://www.reactome.org	Curated knowledgebase of biological pathways. Database of biological processes, covering pathways from basic to high level processes.

2.5 Discussion

Technologies and genomic resources evolve constantly and rapidly. In recent years a very notable increase could be observed in the amounts of information generated and in the methods to analyse it. For example, during the last four years the number of databases increased from approximately 500 available databases to 1,000. Even if there is an interest in a specific

organism, targeting the 'physical' location and available information for the species under study already represents a difficult and time consuming task. Then trying to organize and manage the data to 'extract' or utilize it to give a potential biological meaning to the experiments becomes a great challenge to deal with. The information is dispersed and not well organized to be identified. Additionally, despite the efforts of the main institutes to minimize the inconsistencies between them, like reducing differences between annotated features or the assignment of different identifiers, linking different sources of information remains very complicated. These systems are developed and maintained independently from each other and because of the amounts of data involved, their backend algorithms and standards differ, therefore it is to be expected that there will be differences among the resources.

The use of bioinformatics applications and software to investigate genomic resources has become almost a routine, but with the availability of an extensive list of these applications, selecting the optimal set of tools is difficult and the 'right' assessment of which to utilize for the analyses can consume a considerable amount of time. As a result of this, in most cases it is easier and more typical just to select a 'satisfactory' application.

In addition, several of these tools focus on the analysis of GO enrichment (GO is explained in more detail in Chapter 4). Furthermore, the majority of them apply, or give the option of selecting between different statistical methods to test the enrichment (calculation of p-values, enrichment scores or ratios) and type of correction (i.e. Bonferroni) the application should utilize to perform the analysis. Therefore, the main question is, why are there so many tools to perform the same sort of analysis? This could be because of many reasons: 1) the statistical methods applied are not flexible or cannot be

selected; 2) There is a lack of standardized methods; 3) Visualization and presentation of results vary and the personal preferences to manage the results differ from person to person; 4) It can difficult to find the 'appropriate' application so a new one is developed; 5) The type of identifiers required to make the analyses also plays an important role in the decision of what software should be utilized; If the original gene list is in one format the transformation of a large list to another format can cause several problems: a) it is a time-consuming task, b) some information might be lost (i.e. when the identifier is not found in the 'targeted' identifier type/format), and c) is an additional step that can accumulate errors (i.e. human mistakes, or generated by wrong identifier reference between resources); 6) The performance of the tool (i.e. if it takes too long to load, process the information or give the results, one might choose to try something different), and most importantly; 7) the annotation sources used by the tools and applications.

Basically there is no single tool available that the majority of the research community has found informative, useful and manageable; there is no known tool to meet the requirements of the wider community, so therefore researchers continue creating their own applications or utilizing a satisfactory application.

Although BLAST/BLAT and CLUSTAL have become the standards methods for sequence analysis, and being one of the basic steps in comparative genomics and evolutionary studies, researchers must be careful when drawing conclusions and taking decisions. Wong and colleagues investigated how the choices taken when performing sequences alignment methods for comparative genomics can affect the downstream effects. They used seven different applications to perform sequence alignments and phylogenetic trees for 1,502 sets of orthologous sequences from seven yeast species. The tree

topology results showed that 46.2% of the open reading frames vary in one or more of the resulting trees according to the alignment methods. They also studied other common methodologies that are also dependant on sequence alignments such as substitution rates and the frequency of positively selected sites. The results of substitution rates were not found to be significantly variable among different alignment applications. But, the positive selection analysis identified only 470 ORFs with positively selective sites from which only 44 ORFs were consistent (Wong *et al.* 2008).

Microarray technologies and their methods are far from perfect, and technical and statistical issues still represent a great challenge. Despite this, their application has been successful for recognizing candidate genes that might be responsible or that are contributing to changes on certain phenotypes and/or diseases. The areas of study, where microarrays are applied to, are also extending to other fields such as evolutionary and environmental studies (Gibson 2008). Their application and the flexibility of experimental designs to perform analysis across species may reveal if there is intergenic conservation and open the way to molecular evolutionary analyses (Shiu & Borevitz 2006; Gibson 2008).

Appendix 2.1 Database categories and sub-categories

The total number of databases per class is denoted by the numbers inside the brackets. The complete database list and summaries are available online at the *Nucleic Acids Research* web site (**http://nar.oxfordjournals.org/**).

1. Nucleotide Sequence Databases
    a. International Nucleotide Sequence Database Collaboration (3)
    b. Coding and non-coding DNA (43)
    c. Gene structure, introns and exons, splice sites (24)
    d. Transcriptional regulator sites and transcription factors (64)

2. RNA sequence databases (72)

3. Protein sequence databases
    a. General sequence databases (14)
    b. Protein properties (16)
    c. Protein localization and targeting (23)
    d. Protein sequence motifs and active sites (25)
    e. Protein domain databases; protein classification (38)
    f. Databases of individual protein families (73)

4. Structure Databases
    a. Small molecules (19)
    b. Carbohydrates (9)
    c. Nucleic acid structure (15)
    d. Protein structure (84)

5. Genomics Databases (non-vertebrate) (2)
    a. Genome annotation terms, ontologies and nomenclature (12)
    b. Taxonomy and identification (11)
    c. General genomics databases (45)
    d. Viral genome databases (28)
    e. Prokaryotic genome databases (68)
    f. Unicellular eukaryotes genome databases (19)
    g. Fungal genome databases (31)
    h. Invertebrate genome databases (54)

6. Metabolic and Signaling Pathways
    a. Enzymes and enzyme nomenclature (13)
    b. Metabolic pathways (23)
    c. Protein-protein interactions (77)

    d. Signalling pathways (6)

7. Human and other Vertebrate Genomes (1)
    a. Model organisms, comparative genomics (68)
    b. Human genome databases, maps and viewers (16)
    c. Human ORFs (28)

8. Human Genes and Diseases (1)
    a. General human genetics databases (15)
    b. General polymorphism databases (32)
    c. Cancer gene databases (25)
    d. Gene-, system- or disease-specific databases (56)

9. Microarray Data and other Gene Expression Databases (67)

10. Proteomics Resources (20)

11. Other Molecular Biology Databases (10)
    a. Drugs and drug design (22)
    b. Molecular probes and primers (10)

12. Organelle databases (8)
    a. Mitochondrial genes and proteins (27)

13. Plant databases (1)
    a. General plant databases (42)
    b. Arabidopsis thaliana (27)
    c. Rice (18)
    d. Other plants (18)

14. Immunological databases (27)

Chapter 3 Genetical Genomics

The Genetical genomics concept is extended, putting an emphasis on the experimental designs. Subsequently, a targeted genetical genomics experimental design is presented and analysed. The targeted approach allows the measurement of gene expression levels between two contrasting genotypes of a marked quantitative trait locus. The QTL under study was detected in an inter-cross of broilers and layers on chromosome 4 affecting chicken body weight (Sewalem *et al.* 2002). In this chapter the biological analysis focuses on the interpretation of the annotation related to the 'most' significant genes. This was performed through the manual annotation of the highest differentially expressed genes and the application of curated –mining approaches, which are also introduced briefly in this section. Further re-annotation procedures to obtain homologies, gene ontologies and pathways to a larger set of genes are discussed on Chapter 4 followed by the post-analyses and the investigation of integrative approaches (Chapter 5).

3.1 Genetics and Genomics (Genetical Genomics)

The terminology and idea behind the effort of discovering the mechanisms underlying complex traits by integrating genotypic, molecular, phenotypic and gene expression data was presented in Chapter 1. Typically this technique is referred to as genetical genomics. The potential and applications of genetical genomics have been described in several studies (Morley *et al.* 2004; Monks *et al.* 2004; Kadarmideen *et al.* 2006). However, most of the results to date are somewhat limited because of the current cost for phenotyping (by microarray) all subjects. The cost of gene expression studies

on each individual of the population limits the statistical power for *e*QTL detection. Therefore it is crucial to create experimental designs where a greater statistical significance is achieved (de Koning & Haley 2005).

Genetical genomics has been applied to investigate the mechanisms that contribute to complex trait variation in order to identify genes and pathways underlying a trait or disease, target the downstream effects (i.e. effects caused by gene interactions on molecular pathways), and ultimately to reconstruct and model gene networks. Genetical genomics has also been applied to the identification of molecular subtypes. Using this approach Drake *et al.,* (2006) identified two subgroups of high fat pad mass traits among an F_2 population (B6 x DBA) of mice that appear to be phenotypically similar for fat pad mass.

3.1.1 Experimental Designs

The study of specialized experimental designs can lead us to obtain a greater statistical power at an affordable budget providing the greatest knowledge about complex traits, which can go further than just gene discovery. Researchers must be very careful while selecting and elaborating genetical genomics experimental designs; it has been demonstrated that gene expression results and transcript levels are highly influenced by genetic factors such as cell type, tissue specificity, sequence variations, and heritability (Petretto *et al.* 2006; Arbilly 2006). In addition, Gibson stated that "genotypes are only as good as the environments that they find themselves in", and in order to understand variation underlying complex traits one must not forget to also take into consideration the joint estimation of genotype by environment interactions (Gibson 2008).

Genetical genomics experimental designs require the careful consideration on the number of microarrays needed and the choice of breeds to be used as well as the population design (i.e. F_2, backcross) and the genetic material available. Additionally, because of the cost of phenotyping each individual via microarrays, several selective phenotyping (also referred as treatment choice) techniques have been assessed by only measuring a subset of individuals of the population under study (Jin *et al.* 2004). The number of individuals to be assayed will depend on the microarray platform and the number of slides available. In the case of utilizing a two-channel technology the way of pairing samples, sometimes referred as treatment to unit allocation, also needs to be taken into consideration (Rosa *et al.* 2006).

One approach to combine gene mapping and gene expression data is to study a functional QTL in an experimental cross supplemented by gene expression analysis of the founder lines of the cross. Such studies take the overlap of QTL positions with differentially expressed genes to generate hypotheses about positional candidate genes underlying the QTL (Liu *et al.* 2001; Wayne & McIntyre 2002). However, such designs only study *cis*-acting loci and do not allow one to distinguish the downstream effects of a QTL on expression from effects due to founder line differences unrelated to the target QTL. An experimental strategy to improve the power and efficiency of *e*QTL studies is to focus on one or more known QTL (de Koning *et al.* 2007).

Jansen (2003) developed the method of genetical genomics they presented previously with a technique to refine the original approach by perturbing biological systems in a multifactorial way in the attempt of obtaining a greatest knowledge about complex biological interactions at a lower cost. This technique recommends a detailed evaluation of the experimental designs by choosing the appropriate population for the study, which could

be more informative for complex interactions, or even considering the use of two or more populations for a multifactorial experiment design.

The strategies for selective phenotyping vary from gene dissimilarity techniques to selective transcription. In order to maximize the power of QTL detection Jin *et al.,* (2004) used marker genotype information to select a subset of subjects informative for increased genetic dissimilarity. The algorithm randomly selects the individuals that are homozygous for different alleles, allowing only focusing on the investigation of the additive effects. This genetic dissimilarity selection strategy would not be applicable for studying more complex interactions as they would require the selection of more genotypes. Keller *et al.,* (2005) studied a procedure that would emphasize on the detection of *e*QTLs with dominance effects, however their strategy is not able to differentiate between *cis* (causal) and *trans* (reactive) effects because the scenario they considered was a two inbred lines and their hybrids, as a result only two haplotype configurations are possible for each chromosome. The genetic complementary approach was presented by Bueno Filho *et al.,* (2006). The design of this method is more focused on the goal of the experiment, selecting individuals according to the aim of the study by investigating various scenarios. A very similar approach to selective phenotyping is the selective transcription, where the goal is to maximize the power for detecting associations and correlations between gene expressions and known QTLs by using information from all individuals to perform the analysis (Wang & Nettleton 2006).

Another experimental design factor that could be modified is the generation and appropriate control of biological models by considering the impact of environmental factors on the phenotypes. Borevitz and Chory (2004) suggested that the experimental design of the gene expression study could be

guided by the QTL causing the phenotypic variation, hence the QTL should give some direction to what type of tissue and under what conditions the samples for gene expression studies should be collected. Additionally, the use of a set of conditions in the gene expression study in which the QTL has no effect will provide a control for changes that are not linked to the phenotype (Borevitz & Chory 2004).

Environmental perturbations and genetic plasticity were also used on various experimental designs (i.e. to test gene-by-environment interactions). Landry *et al.*, (2006) investigated genome-wide gene expression of *Saccharomyces cerevisiae* to measure the genetic variation for phenotypic plasticity by studying gene expression of 6 strains on four different environments. They found more than 200 significant genes related to genetic variation for transcriptional plasticity, of which two-thirds showed significant gene-by-environment interaction. The authors argue that the phenotypic plasticity and genotype-by-environment interactions will have important effects on gene-expression network studies as different environments will provide various changes on the ranking of the gene expressions.

A gene expression plasticity experimental design assessing the environmental interaction with *e*QTL on *C. elegans* found differential expression on plastic response to temperature changes with a strong genetic component in the recombinant inbred lines. Interestingly, they were also able to target a common 'master' regulator for a group of 66 co-regulated *trans*-acting genes (Li *et al.* 2006).

As mentioned before, the microarray subject allocations also have a high impact on the design of genetical genomics experiments and therefore

should be optimized. In a single-channel microarray platform, the unit allocation is straightforward. On the contrary, the use of a two-channel microarray platform can considerably complicate the designs. The creation of an intelligent design on a two-channel platform can reduce the number of slides to half of the amount required on a single-channel platform assessing the same number of individuals. The simplest but not as informative design using the two-colour microarrays is the reference design, which is similar to the single-channel layout, although measuring a subject as reference and wasting half of the samples that could be measured. The loop designs are also commonly used on microarray experiments. The basic idea is that it allows the measurement of a subject with both dyes on different slides, meaning that the gene expression of subject x is measured twice, as result the number of subjects assessed would be the same as in a reference design, but with double gene measurements and dye effects accounted for. Chapter 2 explains in more detail commonly used microarray designs.

Fu and Jansen (2006) presented as an optimal design the 'distant pair' approach by selecting individuals with diverse genotype between paired individuals for microarray profiling. The distant pair design pairs the RI strains with the largest genetic difference on the same array, to maximize the amount of useful signal for the QTL mapping. This method could be combined with the selective phenotyping approach when the goal is to estimate additive effects. Recently, an extension of this approach that can be applied to F_2 crosses between outbred lines was presented by Lam *et al.*, (2008).

Furthermore, Li and colleagues presented the 'generalized genetical genomics' approach, which integrates controlled environmental perturbations into a multifactorial experimental design. This approach

investigates the allocation of samples taking into consideration environmental and genotypic information. Through the application of the generalised approach it is expected to detect *e*QTLs across environments, and ultimately to be able to explore heritability and plasticity responses in variable but controlled environments (Li *et al.* 2008).

However, genetical genomics requires a large number of microarrays and its results may be limited by budget constraints. In addition, as the experimental design has a direct impact on the results that will be obtained, it is of great importance to create them according to the goals of the study and the hypotheses to be tested.

3.2 Literature Mining & Data Warehouse

Scientific literature is a major resource that contributes to meaningful gene annotations, although its rapid growth makes it difficult for researchers to keep up-to-date with all the relevant literature and methods related to the field. '-mining' tools have been developed in order to assist the scientists for hypotheses generation and biological discoveries.

Literature-mining tools allow identification, recognition and retrieval of relevant articles through diverse processes. Identification of relevant articles can be achieved by the information retrieval process (e.g. PubMed), which is based on identifying text segments or key words within full articles, abstracts, or sentences. Detection of important biological information cited in the papers can be achieved through the entity recognition process, the goal of which is to find biological entities (e.g. genes, proteins) within the texts. The challenging part of the entity recognition process is due to the lack of

standardization of names; how to recognize all the different and biological names by which the entities are known. The information extraction process aims to detect and extract pre-defined types of facts, like relationships between biological entities (e.g. a *binds* b; c *regulates* a) (Jensen *et al.* 2006).

Literature-mining does not leave much space for novel discoveries, because all the information is based on published articles, although it can be used in genome annotation as curator methods. However, methods such as text-mining tools can lead to novel hypotheses by combining literature-mining tools and information from other sources. Data-mining tools integrate a text-mining approach with other data types (e.g. genome sequences, microarray studies, proteomics) providing a great potential for both approaches (Jensen *et al.* 2006).

Integrated data warehouses are designed to link information between biological networks, sequence analysis methods and experimental results, as well as extending text-mining approaches (Koehler *et al.* 2006). The main problem is how to link all the available information and distinguish the significant data. The process needs to be automated in a systematic and functional way in order to interpret the results. The data warehouse approach can be summarized in the following steps: 1) identification of databases that will be integrated in the framework; 2) extraction of relevant information from the identified sources; 3) processing and conversion of the extracted data, preferably into flat files that should allow the integration into the data warehouse; and, 4) providing access to the data warehouse, which should be efficient (e.g. accessible via internet browser) (Koehler *et al.* 2006; Philippi & Kohler 2006). ArrayExpress is one of the most commonly known and used microarray data warehouse (Parkinson *et al.* 2005). Another good example of this, is ONDEX, a system that aims to integrate databases,

sequence analysis and text mining through graph-based analyses (Koehler *et al.* 2006).

All these '-mining' approaches can face social, technical, and political problems at any stage of their development. Technical challenges can involve, among others, accessibility problems, complicated data extraction, pre-processing data issues, lack of interfaces, erroneous conceptualisation, problems with the content itself of the databases, and error propagation of the sequence annotation. The latter is especially prevalent in the case of automated annotation. Social problems include communication and educational aspects. Licensing and accessibility (free or commercial tools), funding issues and requirements for publications form part of the political difficulties. Some of the suggested solutions are to be very careful while curating the information, provide more documentation of methods and approaches used, provide powerful database interfaces and use of flat files for information retrieval (Koehler *et al.* 2006; Philippi & Kohler 2006).

3.3 Targeted Genetical Genomics

The targeted genetical genomics approach is an experimental design that studies genome-wide gene expression differences for alternative genotypes at a marked QTL segregating in a population. This design concentrates on targeting genetic variation that might be underlying a known QTL. For this thesis, the method was applied to a marked QTL, affecting body weight in chicken, found on chromosome 4 (Sewalem *et al.* 2002). The aim of the study was to identify candidate genes through the effect of the QTL at the gene expression level and to reveal its downstream effects on other genes.

3.3.1 Methods

3.3.1.1 Experimental design

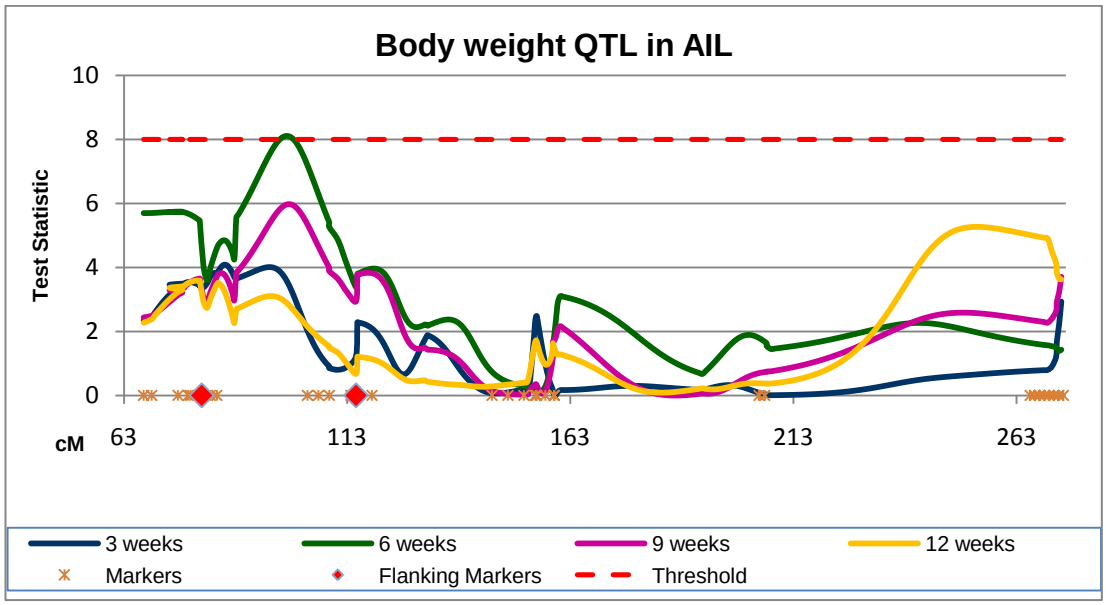
The experimental design consisted of the selection of individuals from an experimental population and the design of the microarray experiment.

The experimental population was initiated from a single F_1 broiler X layer inter-cross family. Twelve F_2 families were mated to produce 60 breeding bird pairs in each generation from the F_2 onwards. The eggs were collected, incubated and pedigree hatched. The chicks were raised in floor pens at controlled diet and light conditions. At 18 weeks of age the adults were caged and fed in limited quantities in order to maintain breeding condition from the F_3. The interval between generations was about 30 weeks. The average inbreeding coefficient in F_7 was 0.26 (with maximum of 0.31). The 372 F_7 progeny was weighted at 3, 6, 9 and 12 weeks. The mean body weights were 386, 1109, 1952 and 2735 g respectively. Blood samples for DNA preparation were obtained at 10 weeks of age.

Individuals from the seventh generation were genotyped for markers covering the entire chromosome 4 (GGA4) (Appendix 3.1). The QTL affecting body weight was confirmed in the AIL, covering a region approximately from 23Mbp to 37Mbp on GGA4 The QTL explains about 5% of phenotypic variation in growth, both expressed as body weight at 6 weeks or growth between 3 and 6 weeks of age (Figure 3.1 ; unpublished data).

The genetic markers snp.28.110.2096.S.1 SNP (rs13576609 NCBI ID; GGA4 position 23,705,378) and snp.3.260.3284.S.2 SNP (rs15544035 NCBI ID; GGA4 position 38,105,938) flanking the QTL were used to infer the QTL genotypes of the birds. The alleles were recorded for all the markers to reflect their line origin (broiler or layer) as derived by the QTL Express software (Seaton *et al.*

2002). Only individuals that were informative and homozygous for the same line origin for the markers flanking the QTL were selected for breeding the birds from the present experiment. It was assumed that the line origin of the broiler corresponded to the *QQ* genotype and those of the layer line corresponded to the *qq* genotype.



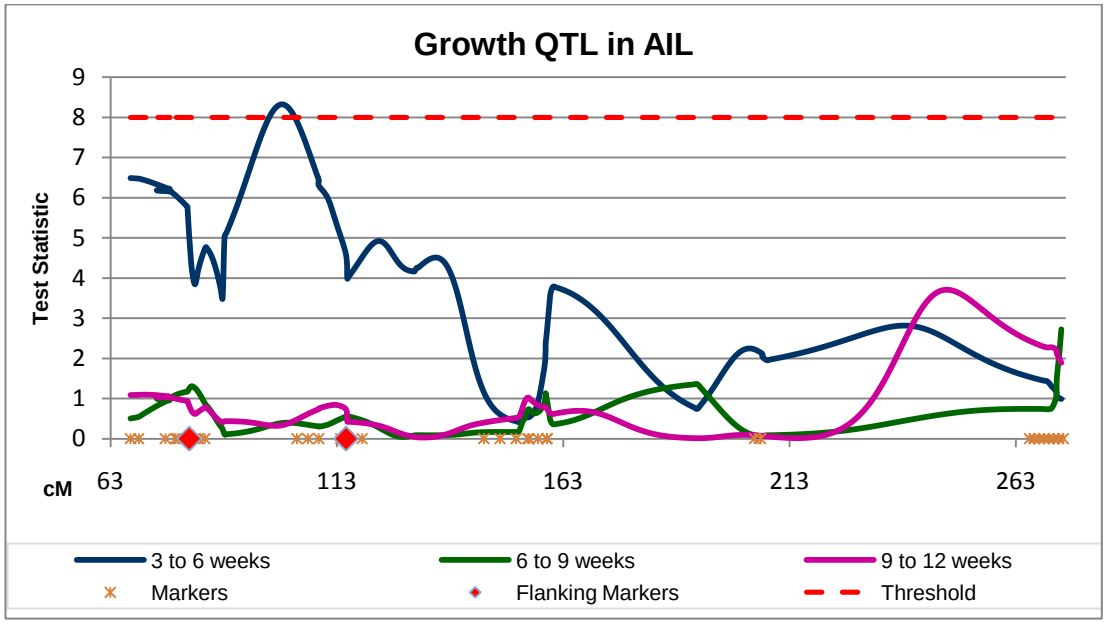


Figure 3.1 QTL graphic for body weight and growth traits on GGA4.

The body weight QTL was confirmed on the AIL at chromosome 4 (approximately from the 80cM to 115cM; ~ 23 Mbp – 38Mbp). The genome wide threshold was set at 8 for both traits (body weight and growth). The markers are represented along the *x* axis, where the red markers represent the flanking markers to infer the QTL genotypes of the birds. Body weights (above frame) and growth (below frame) were recorded for the 372 F2 progeny. Body weights were recorded at 3, 6, 9 and 12 weeks, and growth of the birds was recorded for 3-6 weeks, 6-9 and 9-12.The genome wide threshold is reached by body weight at 6 weeks and growth between 3-6 weeks of age.

Subsequently, breast tissue samples were taken from the progeny at 21 days of age; the samples were taken from eight *QQ* males and 16 *qq* males. The total RNA was isolated using Trizol from the tissue and hybridized onto 16 microarrays to measure gene expression levels (Figure 3.2).

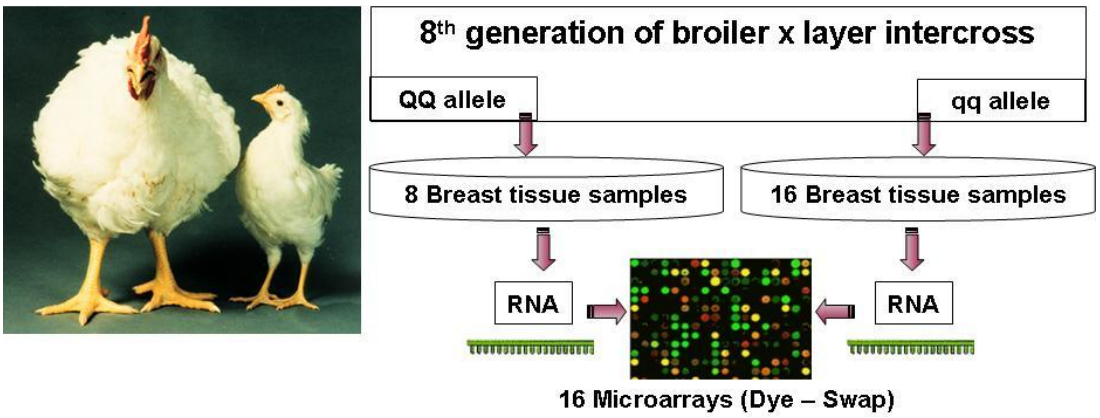


Figure 3.2 Experimental design

Picture: Broiler [left] / Layer [right] at same age) Homozygous individuals for markers flanking a QTL region on GGA4 were identified from the seventh generation of an advanced inter-cross. From the resulting offspring, 8 *QQ* males and 16 *qq* males were slaughtered at 21 days of age; RNA was isolated from breast tissue samples and hybridized onto microarrays to measure the gene expression levels.

The platform for the microarray design was a chicken cDNA array ("ARK-Genomics *G. gallus* 13K v4.0") created from the EST collections generated

with support from the Biotechnology and Biological Sciences Research Council (BBSRC). The array accession identifiers are A-MEXP-831 from ArrayExpress (www.ebi.ac.uk/arrayexpress/); and GPL5673 from the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo/) data repositories. Each array contains 12,877 functional features spotted in duplicate (Burnside et al. 2005). All clones on the array have been sequenced and the information can be found in GenBank at the NCBI server (http://www.ncbi.nlm.nih.gov/nucest/).

A dye-swap design measuring eight pairs of samples twice (a direct comparison of contrasting genotypes) was used (Table 3.1). The microarrays were done by ARK-Genomics, including labelling, hybridization, and image scanning and analysis. Fluorescent labelling of cDNA was performed using the Stratagene Fairplay II microarray labelling kit. The hybridized samples from *QQ* were from individual birds, and each *qq* sample was a pool from two different birds. The microarrays were scanned using the Perkin Elmer ScanArray 5000 according to the manufacturer's instructions. Images were captured at a resolution of 10µm. *Cy*3 images were captured with laser power and PMT gain both set to 80%. *Cy*5 images were captured with the laser power and PMT gain both set to 78% (ARK-Genomics protocol P-MEXP-82356). BlueFuse (© Cambridge bluegnome, 2006) microarray analysis tool was used to quantify the intensities after the slides were scanned.

Table 3.1 Microarray hybridisation design

Slide	*Cy*3 (Green)	*Cy*5 (Red)	*Cy*3-Sample	*Cy*5-Sample
14K 81	*Qq*	*QQ*	qq1.5	QQ3
14K 82	*Qq*	*QQ*	qq1.1	QQ5
14K 83	*Qq*	*QQ*	qq1.7	QQ4
14K 84	*Qq*	*QQ*	qq1.4	QQ6
14K 85	*Qq*	*QQ*	qq1.8	QQ1
14K 86	*Qq*	*QQ*	qq1.6	QQ7
14K 87	*Qq*	*QQ*	qq1.2	QQ8
14K 88	*Qq*	*QQ*	qq1.3	QQ2
14K 89	*QQ*	*Qq*	QQ3	qq1.5
14K 92	*QQ*	*Qq*	QQ5	qq1.1
14K 93	*QQ*	*Qq*	QQ6	qq1.4
14K 94	*QQ*	*Qq*	QQ4	qq1.7
14K 95	*QQ*	*Qq*	QQ1	qq1.8
14K 96	*QQ*	*Qq*	QQ7	qq1.6
14K 97	*QQ*	*Qq*	QQ8	qq1.2
14K 98	*QQ*	*Qq*	QQ2	qq1.3

3.3.2 Analysis

The analysis procedure consisted of: (1) microarray normalization, (2) statistical analysis, and (3) interpretative analyses: a focused analysis under the QTL area, and a global data analysis (downstream QTL effects) of the gene expressions.

3.3.2.1 Microarray normalization

The data were analyzed in R environment using Bioconductor (Gentleman *et al.* 2004). The normalization followed the steps in the Limma package of Bioconductor, with additional plotting from the Marray Bioconductor package. The Limma '*normalizeWhitinArrays*' normalization print-tip loess model was modified. The print-tips consisted in 48 blocks (12 × 4 print-tip arrangement). The control spots were weighted to 0 and the 'data' spots normalized. The normalization of the $M = log\ (Cy5/Cy3)$ values was a 2-step

process of spatial bias correction followed by intensity dependent bias correction. The overall brightness measure of each spot (log-intensity) is defined as $A = (logCy5 + logCy3)/2$. The spatial bias correction was done separately for each block (print-tip), by subtracting corresponding row and column means (excluding control spots) from each "data" spot (Baird *et al.* 2004). The intensity dependent bias was removed by print-tip lowess,

$$N = M - loess_i\ (A)$$

where N = normalized log-ratios (residuals from the tip group lowess regressions; and $loess_i\ (A)$ loess curve as a function of A for the ith tip group. The lowess curve is constructed by performing a series of local regressions (Yang *et al.* 2002; Smyth & Speed 2003).

3.3.2.2 Statistical analysis

The methods followed the Limma package of Bioconductor. A 'swapped' matrix was created with the *M* values multiplied by '-1' for the first slide of each dye-swapped pair. Linear model 'lm.series' analyses were performed on gene means of replicate spots. The design was a "between" stratum defined by pairs of samples, one from each genotype, and the "within" stratum by slides within pairs. The normalized log (*qq*/*QQ*) values were analyzed by separate regression models for each stratum and for each gene. The difference between genotypes was estimated in the between-pairs stratum. The Limma '*eBayes*' correction (Smyth 2004) was used to shrink the residual variances of genes towards their (approximate) median value. The false discovery rate (FDR) was calculated (Benjamini & Hochberg 1995).

3.3.2.3 Interpretative Analysis

The purpose of this analysis was to assign biological meaning to the 'most' differentially expressed genes of the study. Two sets of analyses were

performed: a focused analysis and a global analysis. The focused analysis was designed to investigate expression of genes in the area of the QTL, potentially directly pointing to positional candidate genes for the QTL. The genes were considered as potential positional candidate genes according to their location, level of expression, and biological significance. The global analysis measured the downstream effects of the QTL. A global analysis can identify if the genetic variation was concentrated on particular genetic locations, gene functions, processes, or cellular components. In order to gain the closest biological understanding of the ~ 50 most ('top-analysis') differentially expressed genes, the initial microarray annotation was manually curated and improved.

Initial Microarray Annotation: The initial microarray annotation was provided by ARK-Genomics. This annotation was created by blasting each clone sequence hybridized onto the microarray against various databases (NCBI server [http://www.ncbi.nlm.nih.gov/]; ENSEMBL chicken gene build v36 – December 2005 [http://www.ensembl.org/Gallus_gallus/index.html]; and, IPI (International protein index) from EBI [http://www.ebi.ac.uk/IPI/]). However, usually these types of annotations are on 'static mode', where the data is not automatically updated and any changes made after the annotation was performed will be not included on the data. In order to update and maximize the biological meaning of the experiment, a customized semi-automated re-annotation procedure was performed for all the probes on the array. The semi-automated re-annotation procedure consists of a customized framework following multiple Perl scripts (a commonly used programming language). This procedure uses the outputs produced by the scripts as inputs for other sources (i.e. identifiers re-annotation scripts or input lists for diverse tools) (Chapter 4).

Manual Curated Annotation: Manual annotation was performed on the most differentially expressed genes, in order to have more in-depth and curated data on the significant genes. The initial step was to identify the sources where the data was going to be obtained from. A customized Perl script was used to extract the nucleotide sequences corresponding to the probes immobilised on the array from GenBank (NCBI server [http://www.ncbi.nlm.nih.gov/nucest/]) and searched against the chicken genome sequence and chicken ENSEMBL gene sets [http://www.ensembl.org/Gallus_gallus/index.html Gene Build v36 – December 2005]. In addition, the sequences were investigated using BLAST and compared against human, mouse and rat gene builds in order to identify orthologs. From these various sequence comparisons inferred Gene Ontology annotations, protein family assignments, chromosomal location, descriptions, Interpro and other annotations were accumulated. Annotation was assisted with Data mining [Bio Mart] from ENSEMBL.

3.3.3 Results

3.3.3.1 Animals

Birds selected on their inferred QTL genotype differed significantly in body weight (*QQ* 407.5 ± 63.3 versus *qq* 314.2 ± 52.9; P<0.001 (Mean ± SE)). This corroborates the inferred QTL genotype of the parents of these birds.

3.3.3.2 Microarray Analysis

The microarray contains 12, 877 functional features, which, after annotation, relate to 6,376 unique genes. The raw and normalized *M* values of the microarrays were visualized through print-tip box plots, MA-plots, and spatial plots. These plots were utilized to assess the normalization applied to the chips (Appendix 3.2). Two arrays (14K93_4Cy and 14k94_4Cy) were

considered to perform poorly, although after the normalization procedure and print-tip corrections they were 'comparable' to the other arrays (Figure 3.3). In addition, statistical analyses were also performed by withdrawing the two arrays. The comparison of the 'most' differentially expressed genes list from both analyses did not show significant differences. Therefore, the statistical analyses included all 16 arrays.

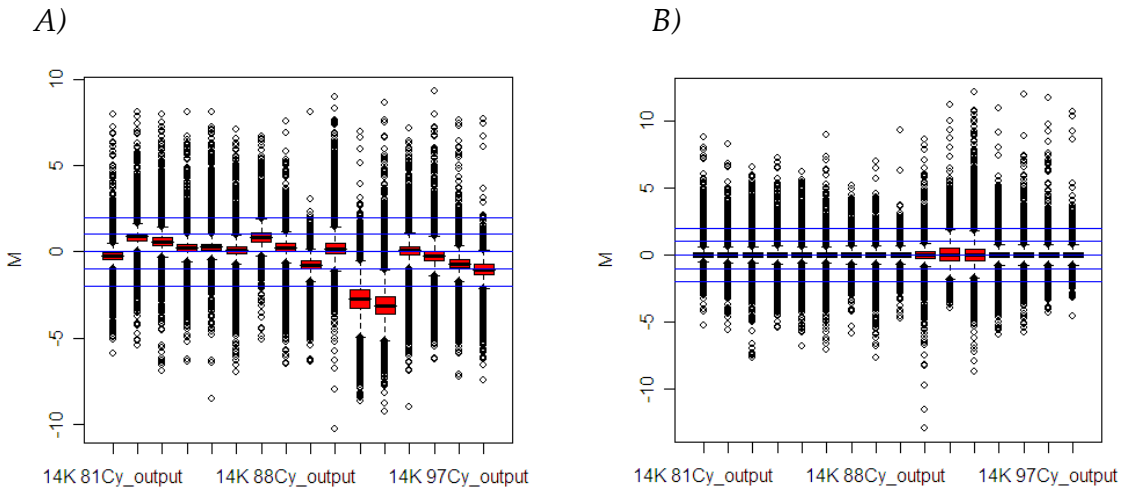


Figure 3.3 Box plots

The box plots represent the interquartile range (25 % - 75%) variability of the 16 microarrays before (*A*) and after (*B*) normalization, with the median represented with the horizontal black line. (A)Box plots of the raw *M*-values. Data values were not centred on 0 and two arrays (11-14K_93 and 12-14K_94) display a larger variance than the others and their intensities also greater than the range of the other arrays; (B) Box plots of the normalized *M*-values, the values were centred on 0 and the variances between and within chips was reduced.

The statistical analysis of the microarray data found 580 transcripts differentially expressed between the two genotypes at an FDR threshold level of 30%. With the application of this threshold, about 174 genes among the 580 genes are expected to be false positives. From the differentially expressed genes, 315 transcripts (186 unique genes) were up-regulated in the

muscle of birds with *qq* genotype and 265 (172 unique genes) were up-regulated in birds with *QQ* genotype.

Most of the differentially expressed genes, map outside the region of the QTL under study, suggesting the QTL has many *trans*-acting effects (Figure 3.4). However, four differentially expressed genes were located within the QTL region (~23 - 37 mbp) and for these the QTL can be considered as *cis*-acting, making these genes positional candidates for further studies.

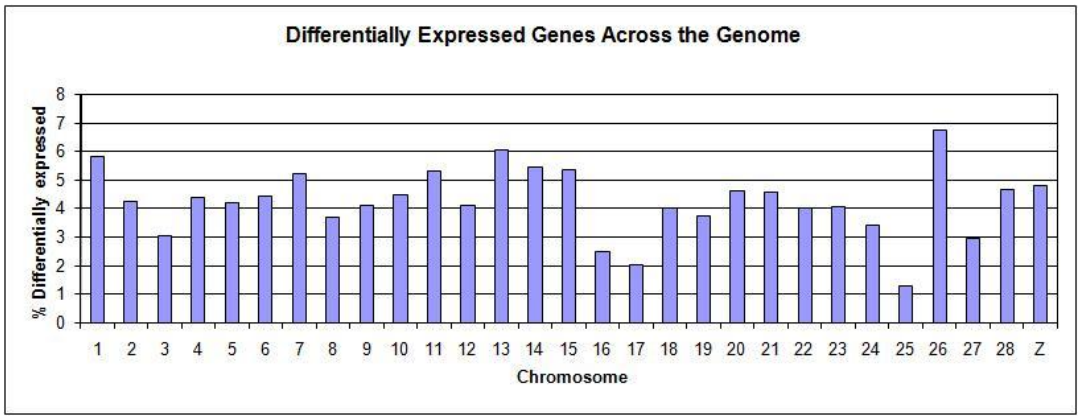


Figure 3.4 Differentially expressed genes across the genome.

Differentially expressed genes across the chicken genome, given as the percentage of genes from the total genes per chromosome present on the microarray. Some chromosomes (i.e. chromosome 26) show a high percentage of differentially expressed genes due to the low number of genes annotated to the chromosome and the physical size of the chromosome.

Focused analysis under the QTL: Chicken chromosome 4 is approximately 94 Mbp long, containing around 1,187 genes (from which ~ 498 genes are represented on the microarray). Figure 3.5 represents the position on GGA4 of transcripts on the microarray with their respective *t*-value for the contrast between *QQ* and *qq*.

The confidence interval of the QTL under study covers almost 14% of the chromosome 4 (23 Mbp ~ 37 Mbp). This region, according to present

annotation, contains around 139 genes, from which approximately 58 genes are present on the microarray. In total, there were four differentially expressed clones in the region of the QTL (BU452163; BU415609; BU465968; and BU463895) but they were very poorly annotated. The biological functionality of these positional candidate genes led to further studies on downstream effects and genome variation as described in the global analysis.

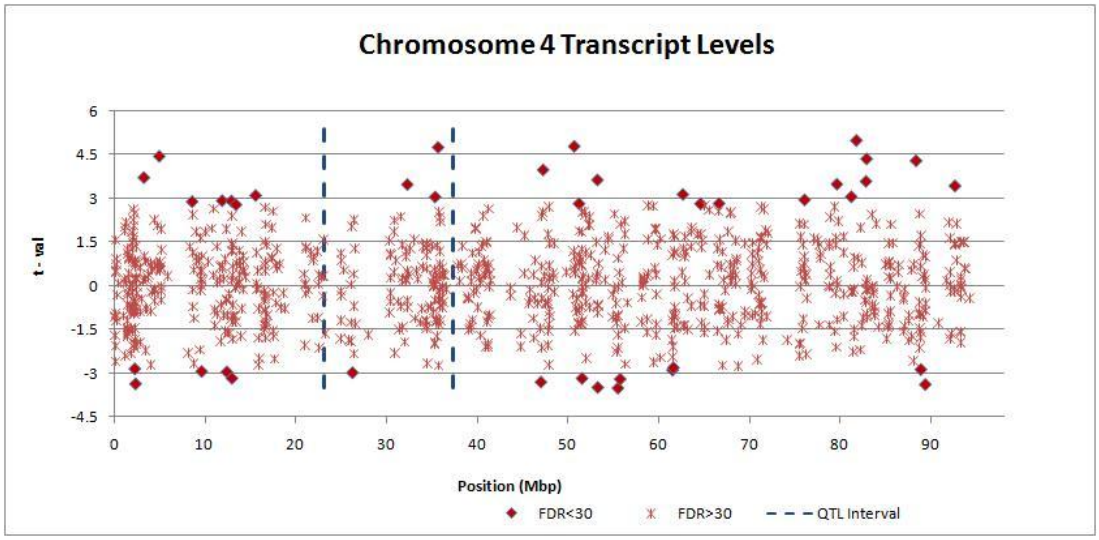


Figure 3.5 Chromosome 4 transcripts levels

Distribution of transcripts according to their position in mega base pairs across GGA4, with their respective *t*-value for each transcript on the *y*-axis. Negative *t*-values represent transcripts that are up-regulated for the *QQ* genotype. Positive *t*-values represent transcripts that are up-regulated for the *qq* genotype. The QTL region is by denoted the horizontal dotted lines.

Global analysis Figure 3.4 represents the distribution of the differentially expressed genes across the chicken genome given as percentage of genes on the array from each chromosome. Multiple biological processes, cellular components and molecular functions were linked to the differentially

expressed genes (e.g. DNA replication initiation, regulation of transcription, zinc ion binding, chromatin).

Data mining on the top 50 differentially expressed transcripts identified several genes that are potentially involved in body size, muscle development and skeletal muscle differentiation (Table 3.2; more detailed description in Appendix 3.3). Examples of these genes that were up-regulated for the *QQ* genotype were *superoxide dismutase 1* (SOD1), *dachshund homolog 1* (DACH1), *zinc finger protein 367* (ZNF367), *purinergic receptor P2Y G-protein coupled 5* (P2RY5), *oxysterol binding protein-like 6* (OSBPL6), and *lectin-like protein type II (17.5)*. A preponderance of genes up-regulated for the *qq* genotype involved in pathway inhibition was observed (i.e. *msh homeo box homolog 1*, MSX1; *leptin receptor overlapping transcript-like 1*, LEPROTL1).

In summary, there was no particular enrichment in the number of differentially expressed genes under the QTL region, however this was not expected because genes affecting the same phenotype might not be clustering on the same region. The focused analyses allowed us to identify four positional candidate loci, although only three of them could be characterized in terms of gene function. The global analyses presented consistent functional results, and detected what might be downstream effects of the QTL. These results will be discussed in more detail in Chapters 4 and 5.

Table 3.2 Top 40 differentially expressed genes and their annotation

Genbank	*FDR*	*p – value*	*Gene*	*Chr*	*Description*
BU239793	0.037	4.02E-06			similar to C1 inhibitor precursor
BU464781	0.135	5.98E-05	ORC2L	7	similar to Orc2-A-prov protein
BU403621	0.153	6.61E-05	KSR1	19	Clone ChEST370a4

BU449186	0.185	7.27E-05			homologue to Nhn1A
BU315042	0.212	1.69E-04	ZNF367	Z	zinc finger protein 367
BU306178	0.214	2.00E-04			Probable cation-transporting ATPase 13A
BU141564	0.220	2.38E-04	MYO3B	7	similar to Myosin IIIB
AI981966	0.222	2.79E-04	P2RY5	1	P2Y purinoceptor 5
BU325342	0.223	2.94E-04	ANAPC13	9	Anaphase promoting complex subunit 13
BU322163	0.225	2.95E-04			similar to transcription factor AP-2 beta-like 1
BU410005	0.226	3.20E-04	OSBPL6	7	oxysterol binding protein-like 6
BU404735	0.227	3.67E-04	LOC416512	14	homologue to MGC3048 protein
AI981398	0.229	3.89E-04			similar to Envelope protein
BU121813	0.230	4.07E-04	ITPR2	1	homologue to Inositol trisphosphate receptor type 2,
BU262117	0.230	4.20E-04			Similar to tRNA isopentenyltransferase 1
BU242249	0.230	4.88E-04	LOC416715	3	similar to brain-selective and closely mapped on the counter allele of CMAP in cystatin cluster
BG625280	0.230	4.98E-04	LOC776458	Un	Lectin-like protein; type II transmembrane protein (17.5)
BU204092	0.231	5.21E-04	MSX1	4	homeobox protein GHOX-7
BU226228	0.231	5.84E-04	DACH1	1	Similar to dachshund 1
BU346776	0.234	6.24E-04	ZFPM2	2	Similar to Zinc finger protein ZFPM2
DN829950	0.235	7.06E-04			Leptin receptor overlapping transcript-like 1
BU359115	0.236	7.14E-04			homologue to Flj38101-prov protein, partial (98%)
BU289016	0.236	7.20E-04	LOC771059	27	
BU200856	0.241	7.92E-04	SHQ1	12	SHQ1 homolog (S. cerevisiae)
BU375548	0.242	8.60E-04	ADRA2B	13	Similar to alpha2Da-adrenoceptor
BU262776	0.242	9.03E-04	RCJMB04_5h22	20	Death inducer-obliterator 1
BU386891	0.242	9.15E-04	RCJMB04_5f7	1	Similar to Histone-lysine N-methyltransferase
BU352910	0.242	0.0009621	MYLIP	2	Myosin regulatory light chain interacting protein
BU259214	0.242	1.01E-03	ASB14	12	similar to ankyrin repeat and SOCS box-containing protein 14
BU467727	0.243	1.02E-03	VPS54	3	
BU468014	0.245	1.02E-03			similar to selenoprotein M precursor (*Homo sapiens*)
BU470314	0.246	1.04E-03	COQ5	15	ubiquinone biosynthesis methyltransferase activity
BU466922	0.247	1.18E-03	ITPKB	3	Similar to Inositol-trisphosphate 3-kinase B
BU142536	0.247	1.20E-03	F9	4	coagulation factor IX
AI981000	0.252	1.20E-03	P2RY5	1	Variant retinoblastoma transcription factor
BG625071	0.254	1.22E-03	COL4A1	1	collagen, type IV, alpha 1
BU361603	0.257	1.23E-03	SOD1	1	superoxide dismutase 1
AJ399183	0.258	1.26E-03	ATM	1	ATM (ataxia telangiectasia mutated)
BF723876	0.259	1.38E-03	TBC1D14	4	TBC1 domain family, member 14

3.3.4 Discussion

With the application of the experimental design of targeted genetical genomics we attempted to narrow the gap between QTL and phenotype. Targeted genetical genomics allows the direct quantification of the link between the genotypes and the genetic responses. As mentioned previously, an important assumption made in typical genetical genomics studies is that the gene expression values are also affected by what is causing the differences on the traits. An advantage of the 'targeted' study is that the body weight QTL region has been mapped previously; and that the QTL was also confirmed on the population of this study, from which, birds with alternative genotypes were used for expression profiling. Therefore, the gene expression changes are expected to reflect the variability between the genotypes.

Targeted genetical genomics facilitates the prioritization of potential positional candidate genes responsible for the QTL by linking the location of the genes, their expression and the biological impact of the QTL. Four candidate genes and functional annotations related to a body weight QTL on chicken chromosome 4 were identified. Additionally, the application of this method is financially attractive because of the nature of the experimental design and the amount of microarrays required for the investigation of the genetic variation of a known QTL. However, a similar experimental design with a greater number of samples, and if possible with a time-course and various tissue hybridizations, would allow a more profound investigation, where clustering techniques and possibly network modelling could be applied (Emilsson *et al.* 2008).

The four significant genes under the QTL region were classified as potential positional candidate genes. Unfortunately, the initial annotation of these genes is not informative, and in order to obtain a better characterization of

these genes further in-depth analyses are required (Chapter 5). Additionally, is important to note that not all the genes under the QTL region are linked to gene expression levels (not hybridized onto the microarrays), resulting in potential positional candidate genes being undetected.

The functional study of the top differentially expressed genes recognized various genes that are involved in body size, muscle development and skeletal muscle differentiation (Appendix 3.3). Some of these results are discussed in more detail: SOD1 was linked to numerous mouse gene ontology terms such as regulation of body size, muscle maintenance, activation of MAPK activity. DACH1 was related to regulation of transcription, and may inhibit *transforming growth factor β* (TGF-β) (Wu *et al.* 2003). ZFF29b, a splice variant of ZNF367, was identified as a potential activator of erythroid gene promoters (Asano *et al.* 2004). The *lectin-like protein* was found three times on the top 50 significant genes, and contains a *C-type lectin-like domain* (CTLD) found in natural killer cell receptors (Marchler-Bauer *et al.* 2007). Lectins contain discrete carbohydrate-recognition domains which regulate several complex carbohydrate biological effects (Drickamer & Fadden 2002).

The *msh homeo box homolog 1* (MSX1) is a highly differentially expressed gene (*p*-value 5.21E-04) located on GGA4; this gene is involved in pathway inhibition and it is up-regulated for the *qq* genotype. MSX1 cooperates with histone (H1b) to inhibit transcription and myogenesis and as with other MSX homeoproteins inhibits the differentiation of skeletal muscle, acting as a negative regulator of muscle differentiation (Woloshin *et al.* 1995; Lee *et al.* 2004).

Leptin is associated with obesity in other species, and it decreases food intake when infused in chickens (Rosenbaum & Leibel 1998; Kuo *et al.* 2005). However the existence of the *leptin* gene is controversial in chickens (Friedman-Einat *et al.* 1999; Sharp *et al.* 2008). Although the *leptin* ligand has not been found, a paralogous gene to the *leptin receptor* (LEPROTL1) was found on GGA4 also outside the QTL region and showed a significant difference in gene expression between the QTL genotypes.

As in many other approaches, the potential of this method can be limited by the level of genome annotation and the bioinformatics resources of the organism under study. In chicken, the present annotation is reasonable, but is not as complete and informative as other model organisms. In spite of the availability of the chicken genome sequence, some clones (~ 200 of the differentially expressed transcripts) on the initial annotation could not be assigned to their genome allocation. Additionally, approximately only 41% of the genes under the QTL region were represented on the microarray. However, the use of bioinformatics methods and comparative mapping approaches can improve the annotation of the genes present on the microarray, allowing the identification of unnoticed gene functions when using chicken-specific data only. Although targeted genetical genomics was applied to chicken, this approach and methodology could be applied to other species. On the other hand, the chicken was a very useful organism for this study, because the trait under study is commercially, scientifically and medically important. These methods and results can help us to identify and/or discover genetic mechanisms underlying a particular trait affected by a QTL.

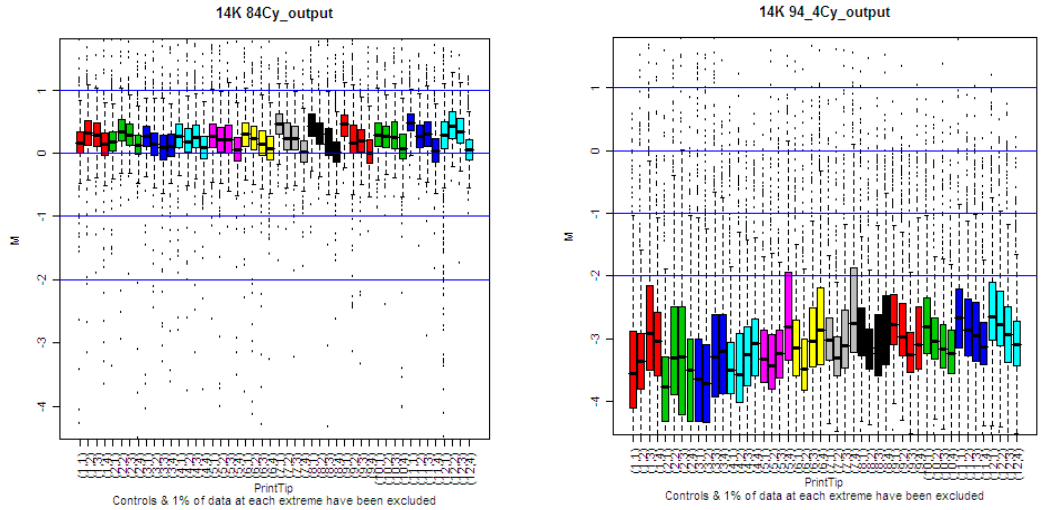
Appendix 3.1 Markers across GGA4

Marker	Position at Chr4	cM on GGA4
snp_11_34_12059_S_1	17,417,516	67.4
snp_272_1_2569_S_1	19,944,735	69.2
snp_28_9_1253_S_2	20,460,191	75.1
snp_28_18_109662_S_1	20,989,073	77.4
snp_28_60_21850_S_2	22,021,512	77.9
snp_28_110_2096_S_1	23,188,226	80.4
snp_28_151_7417_s_1	23,857,180	82.4
snp_28_197_8877_S_2	25,018,997	83.8
snp_28_273_40501_S_2	27,293,510	104.1
snp_28_388_22709_s_1	30,051,093	106.6
snp_3_134_131740_S_1	34,776,111	109.0
snp_3_167_49373_s_1	35,341,191	113.9
snp_3_198_28795_S_2	36,331,875	114.4
snp_3_260_3284_S_2	37,392,928	115.0
snp_3_538_25371_s_1	43,633,356	118.6
snp_3_627_80329_s_1	46,075,077	145.5
snp_3_718_65599_S_1	48,374,000	149.0
snp_3_824_29936_S_1	51,211,698	152.6
snp_3_834_44887_S_1	51,462,545	155.2
snp_3_857_41571_S_1	52,470,252	155.4
snp_3_886_45694_S_2	53,227,816	157.4
snp_3_890_70076_S_1	53,494,953	159.4
snp_3_897_86144_S_2	53,748,987	159.4
snp_32_183_40414_s_1	63,792,065	205.3
snp_32_90_6466_S_1	66,768,365	206.6
snp_31_135_18627_s_1	80,706,534	266.0
snp_31_121_14943_s_1	81,461,219	267.1
snp_31_121_14845_s_1	81,461,317	268.1
snp_31_121_14537_s_1	81,461,625	269.2
snp_31_121_14506_s_1	81,461,656	270.3
snp_31_48_37458_s_1	83,334,213	271.3
snp_31_42_5331_S_1	83,578,827	272.4
snp_31_22_42603_S_2	84,085,739	273.5

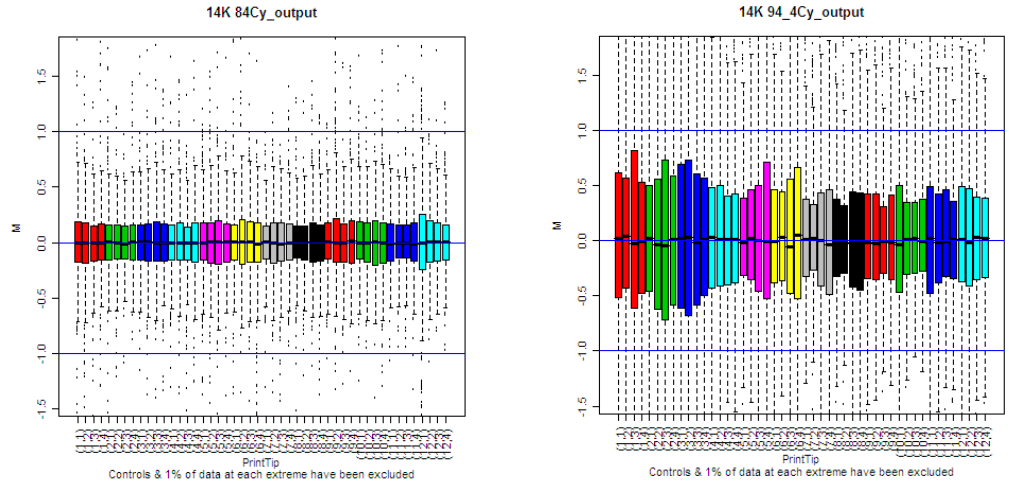
Appendix 3.2 Raw and Normalized microarray plots

Two microarrays are used to display various visualizations (A: print-tip box plots; B: M-A plots; C: Spatial plots). The chip 14K 84Cy (left) is used to illustrate the chips with 'normal/good' quality arrays. And, 14K 94_4Cy array (right), illustrates one 'low-bad' chip.

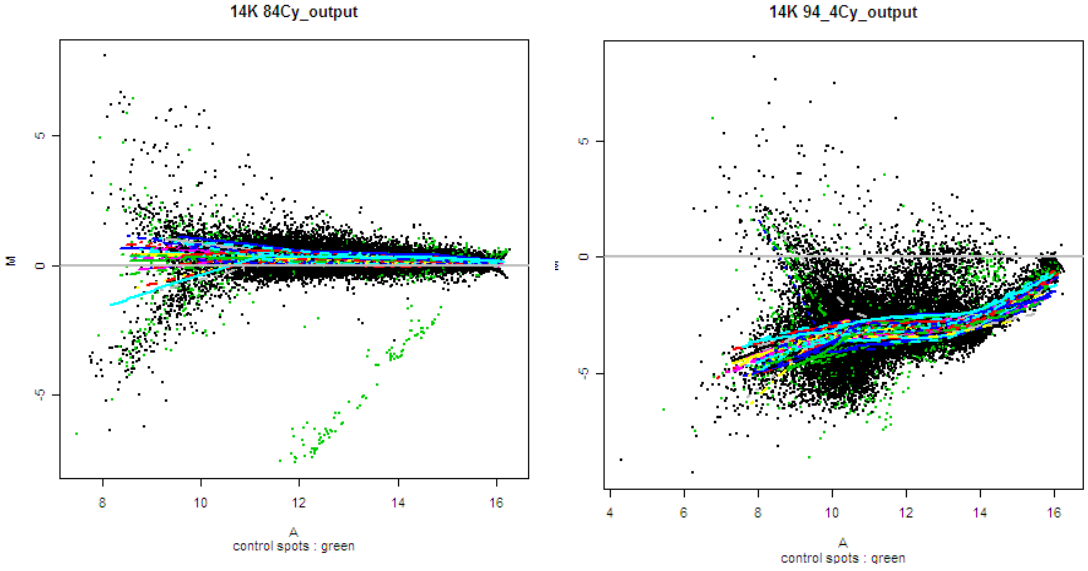
A) Print-tip box plots of M values of Raw data: Distributions from 48 print-tips (12 × 4), none is centred at 0. Left plot (14K 84Cy) demonstrates a small print-tip effect (more visible from the 7th (grey) – 12th (aqua) print-tip groups. Right plot (14K 94_4Cy) distributions are further from 0 and the print-tip variances are larger.



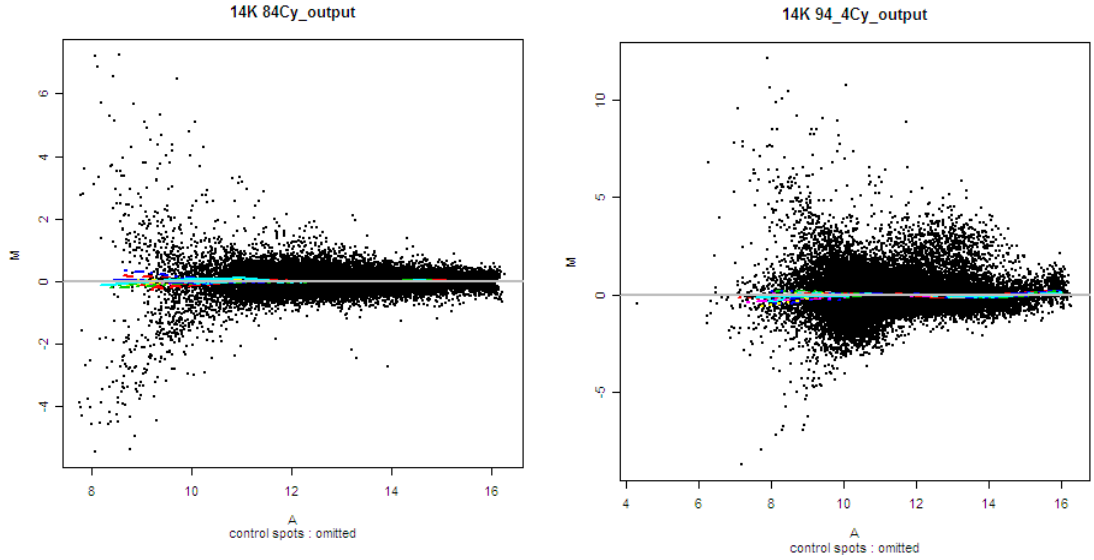
Print-tip box plots of M values of Normalised data: Distributions from 48 print-tips (12 × 4) after normalization. Distributions are centred at 0. Array 14K 84Cy demonstrates that the print-tip effect was removed. 14K 94_4Cy was slightly corrected although high variance can be observed.



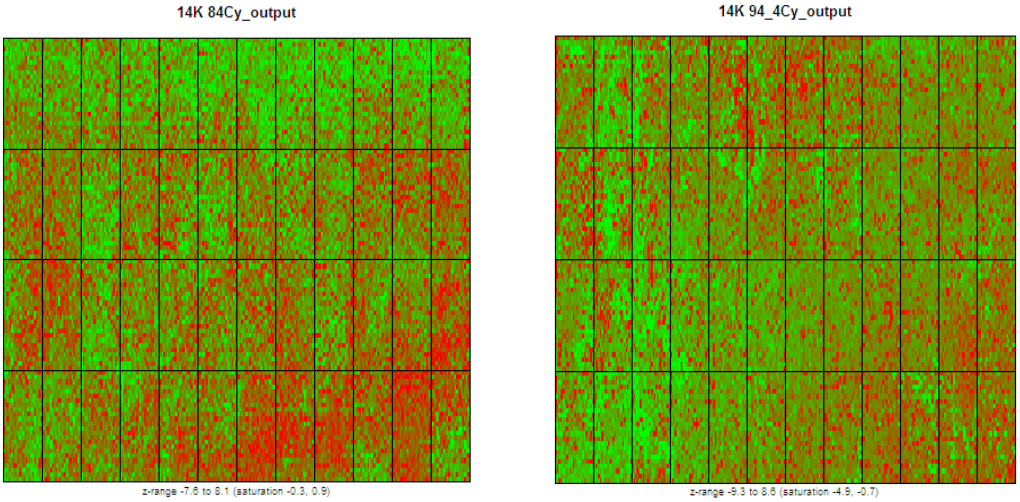
B) M-A Plots of Raw Data: Un-normalized values of *Mjk = log (Cy5jk/Cy3jk)*, where *k* is the gene on array *j;* and, *A = ((logCy5jk) + (logCy3jk))/2*. The plots include control spots (green spots). Coloured lines represent print-tip intensity dependent trends modelled by a regression curve. Array 14K 94_4Cy shows non-uniform intensity trend.



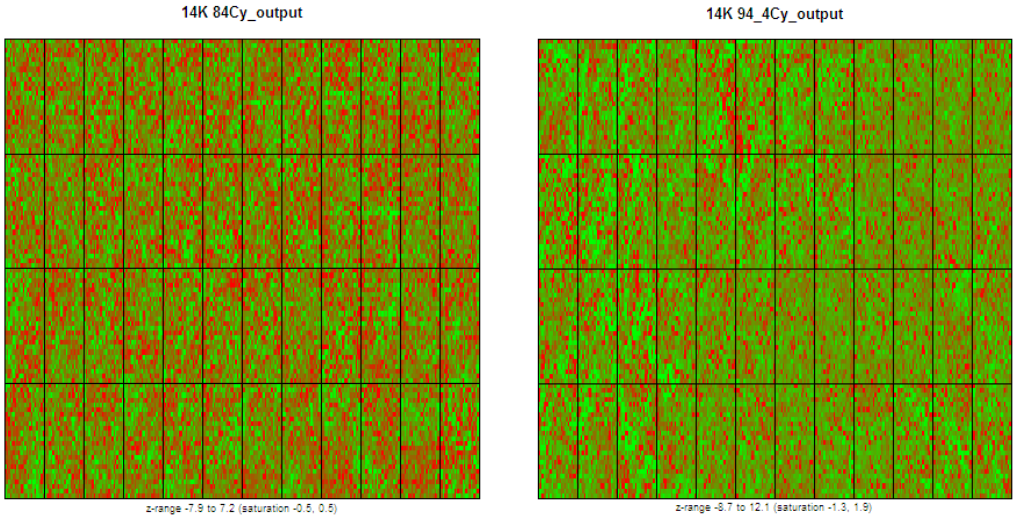
M-A Plots of Normalized Data: normalized log-ratios (residuals from the tip group loess regressions. The correction of a 'low-bad' quality array (14k94_4Cy) was improved and comparable to the other arrays of the study.



C) Spatial Plots of Raw data (heat-maps): The data is represented by 48 blocks (print-tips 12 × 4 arrangement). Array 14K 84Cy shows at least two regions where the intensities could be due technical artefacts. One region was noticed in the print-tips of the first row (1.1 -12.1) where the green brightness is more predominant than the red one. Other visible sections are the print tips 7.4 – 9.4 with more 'red' intensities observed. Array 14k94_4C tends to have an overall 'green' intensity.



Spatial Plots of Normalized data (heat-maps): Spatial artefacts observed (above) were removed after normalization.



Appendix 3.3 Detailed Annotation

Top 50 differentially expressed genes detailed annotation. (Including only those where functional information could be related to the gene)

GenBank	Gene	*t*-value	*p*-value	Chr	Description	Related term(s)
BU231026	TCTN2	-6.56	5.6E-05		Tectonic-2 precursor	Hedgehog (Hh) pathway activation and repression
BU403621	KSR1	-6.44	6.6E-05	19	Kinase suppressor of Ras 1	ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding, Serine/threonine protein kinase, active site
BU315042	ZNF367	-5.74	1.7E-04	Z	Zinc finger protein 367	Potential activator of erythroid gene promoters, catalytic activity, nucleic acid binding, zinc ion binding
AI981966	P2RY5	-5.39	2.8E-04	1	Purinergic receptor P2Y, G-protein coupled, 5	Variant retinoblastoma transcription factor, purinergic nucleotide receptor activity, G-protein coupled, receptor activity, rhodopsin-like receptor activity
BU325342	ANAPC13	-5.35	2.9E-04	9	homologue to Anaphase promoting complex subunit 13	Involved in mitosis
BU410005	OSBPL6	-5.30	3.2E-04	7	Oxysterol binding protein-like 6	Pleckstrin-like domain
BU404735	LOC416512	-5.20	3.7E-04	14	homologue to MGC3048 protein, partial (84%)	
AI981398		-5.16	3.9E-04		LOC772243 similar to gag/env fusion protein	aspartic-type endopeptidase activity, nucleic acid binding, structural molecule activity, zinc ion binding
BU242249	LOC416715	-5.01	4.9E-04	3	Similar to brain-selective and closely mapped on the counter allele of CMAP in cystatin cluster	Catalytic activity, Soluble quinoprotein glucose dehydrogenase, Strictosidine synthase
BG625280	LOC776458	-5.00	5.0E-04	Un	Lectin-like protein; type II transmembrane protein (17.5)	C-type lectin-like domain found in natural killer cell receptors; sugar binding
BU226228	DACH1	-4.89	5.8E-04	1	Dachshund homolog 1	Regulation of transcription, may inhibit tranforming growth factor ß, Putative DNA binding, Protein kinase core, Transforming protein Ski, Kv1.4 voltage-gated K+ channel, Antifreeze protein, type I, Gamma-glutamyltranspeptidase

BU386891	RCJMB04_5f7	-4.60	9.1E-04	1	Similar to Histone-lysine N-methyltransferase, H3 lysine-9 specific 2 (Histone H3-K9 methyltransferase 2) (H3-K9-HMTase 2) (Suppressor of variegation 3-9 homolog 2) (Su(var)3-9 homolog 2)	chromatin binding,histone-lysine N-methyltransferase activity,protein methyltransferase activity , transferase activity, zinc ion binding, SET domain, Pre-SET zinc-binding sub-group, Post-SET zinc-binding region, Histone H3-K9 methyltransferase
BU259214	ASB14	-4.54	1.0E-03	12	Similar to ankyrin repeat and SOCS box-containing protein 14	intracellular signaling cascade
BU470314	COQ5	-4.52	1.0E-03	15	Coenzyme Q5 homolog, methyltransferase (S. cerevisiae)	methyltransferase activity, transferase activity
AI981000	P2RY5	-4.43	1.2E-03	1	Variant retinoblastoma transcription factor	
BG625071	COL4A1	-4.42	1.2E-03	1	Collagen, type IV, alpha 1	binding, extracellular matrix structural constituent
BU361603	SOD1	-4.41	1.2E-03	1	Superoxide dismutase (Cu-Zn)	Regulation of body size, muscle maintenance, activation of MAPK activity
BF723876	TBC1D14	4.34	1.4E-03	4	TBC1 domain family, member 14	
AJ399183	ATM	4.40	1.3E-03	1	Ataxia telangiectasia mutated (includes complementation groups A, C and D)	
BU142536	F9	4.43	1.2E-03	4	Coagulation factor IX , partial (65%)	calcium ion binding, coagulation factor IXa activity, peptidase activity, Involved in epidermal growth factor
BU467727	VPS54	4.53	1.0E-03	3	Vacuolar protein sorting-associated protein 54	protein binding, Domain bacterial regulatory factor, effector; Vps54-like
BU352910	MYLIP	4.57	9.6E-04	2	Myosin regulatory light chain interacting protein	cytoskeletal protein binding, metal ion binding, zinc ion binding, Ezrin/radixin/moesin ERM domain,Band 4.1, N-terminal
BU262776	DIDO1	4.61	9.0E-04	20	Death inducer-obliterator 1	metal ion binding,protein binding,zinc ion binding,spen paralogue and orthologue C-terminal,Proline-rich region, Zinc finger
BU375548	ADRA2B	4.64	8.6E-04	13	Similar to Alpha2 d2 adrenergic receptor	adrenoceptor activity, rhodopsin-like receptor activity, sodium:dicarboxylate symporter activity ,5-Hydroxytryptamine 6 receptor, 5-Hydroxytryptamine receptor
DN829950		4.77	7.1E-04		Leptin receptor overlapping transcript-like 1	Vacuolar protein sorting 55
BU346776	ZFPM2	4.85	6.2E-04	2	Zinc finger protein	Zinc ion binding
BU204092	MSX1	4.97	5.2E-04	4	Homeobox protein GHOX-7 (CHOX-7) (Hox-7)	Cooperates with H1 to inhibit transcription and myogenesis, regulation of transcription, DNA-dependent, multicellular organismal development

BU121813	ITPR2	5.13	4.1E-04	1	Inositol 1,4,5-triphosphate receptor, type 2	calcium channel activity, inositol 1,4,5-triphosphate-sensitive calcium-release channel activity, ion channel activity, RyR and IP3R Homology associated (Ryanodine receptor activation is a key component of muscular contraction, their activation allowing release of Ca2+ from the sarcoplasmic reticulum. Mutations in the ryanodine receptor lead to malignant hyperthermia susceptibility the and central core disease of muscle)
BU141564	MYO3B	5.50	2.4E-04	7	Myosin-IIIB	Myosin head, motor region, IQ calmodulin-binding region major calcium sensor and orchestrator of regulatory events
BU449186	NP_653205.2	6.37	7.3E-05		Conserved nuclear protein NHN1	Zinc finger, CCCH-type, Tubby N-terminal (related to obesity in mice), Proline-rich region
BU464781	ORC2L	6.52	6.0E-05	7	Origin recognition complex, subunit 2-like (yeast)	DNA replication initiation
BU239793		8.85	4.0E-06		Similar to MGC11257 protein, (6%)	

Chapter 4 Annotation Procedures

4.1 Introduction

The objective of annotating genes in high-throughput experiments goes further than only relating them to biological terms. The annotation also attempts to understand how the genes work together in order to manifest changes on certain traits or diseases. This requires the investigation of processes to identify genes and subsequently finding the pathways, gene ontologies, functions and procedures in which they are involved.

One of the main difficulties of annotating the data is to effectively manage the multiple resources where the information can be obtained from. Annotations can be derived from diverse sources (i.e. sequence analyses and experimental results) and typically are classified as structural and functional. Basically, the structural annotation refers to the assembly of the genome and the functional annotation refers to the 'roles' of the genes and their products (McCarthy 2007). Genetic data are spread over several databases and usually developed independently, complicating the integration and use of these resources.

Furthermore, the availability of so many resources requires a systematic approach to be able to deal with the information and exploit the data to its maximum potential. The lack of annotation represents one of the major limiting factors for the interpretation of results from high-throughput experiments. Therefore, the research to improve annotation tools and their integration methods is of great importance. In this chapter, a customized, semi-automated annotation pipeline developed for the annotation of the

probes of the ARK-Genomics *G. gallus* 13K v4.0 cDNA array is presented. The microarray original annotation (Chapter 3) and all the information obtained through the framework were centralized on a project-specific relational database, in order to be able to perform in-depth biological and functional analysis. The post-analyses performed on the data are illustrated in Chapter 5, considering the investigation of methods for functional analyses such as conservation across species, protein domain, and alternative splicing sites.

4.2 Ontologies

There is a huge research community working on genetics, bioinformatics, medicine, molecular biology and related fields which can benefit from integrating and comparing their findings. However, it is very difficult to use the exact same terminology and definitions across scientific fields or even across research groups. The use of ontologies (controlled standardized vocabularies) aims to reduce the variation among the terms and annotations utilized, and might allow scientists to make data comparable and manageable for computational analysis. The success of the creation and application of ontologies led to their proliferation, creating an 'extra' obstacle for data integration. The Open Biomedical Ontologies (OBO; http://www.obofoundry.org/) foundry aims to coordinate the reform, standardization and creation of interoperable biological ontologies (Smith *et al.* 2007).

The gene ontology (GO) consortium is one of the most successfully applied and developed ontologies which describes gene and gene product attributes. The terms are independent of the species and are associated mainly in three

categories: biological processes, cellular components, and molecular functions. When the genes and gene products are associated with several terms, they are annotated on each of these terms. Also, because of the nature of the gene ontology structure, genes that are annotated to a given term are also annotated to all its 'parental' terms.

Within the OBO Foundry, we can find more examples of ontologies such as PRotein Ontology (PRO; http://pir.georgetown.edu/pro/) which describes properties attributable to proteins and their relationship to GOs (Natale *et al.* 2007); and also Plant Ontology Consortium (PO; http://www.plantontology.org/) dedicated to control and curate plant growth, structure and development vocabularies (Pankaj Jaiswal *et al.* 2005), among others.

Current gene ontology statistics (as of January 25 of 2009) are: 26,647 terms, 98.5% with definitions; 15,900 related to a biological process; 2,256 related to cellular component; and 8,491 related to molecular function (http://www.geneontology.org/).

Gene ontology annotations are not strictly assigned by curators; they are also annotated through automated processes. GO terms include an evidence code to specify how the particular term is supported. The purpose of the code is not to determine the quality or type of work, but should be used in conjunction with the analyses performed. Evidence codes are divided in four categories: 1) experimental (cited articles displaying physical characterization of gene and/or gene products), 2) computational analysis (i.e. based on an *in-silico* sequence analysis), 3) author statements and 4) curatorial statements (Table 4.1 Gene ontology evidence codes). The only evidence code that is not

assigned by a curator is the one denoted as IEA (Inferred from Electronic Annotation).

Table 4.1 Gene Ontology Evidence Codes

Category	Abbreviation	Definition
Experimental	EXP	Inferred from Experiment
Experimental	IDA	Inferred from Direct Assay
Experimental	IPI	Inferred from Physical Interaction
Experimental	IMP	Inferred from Mutant Phenotype
Experimental	IGI	Inferred from Genetic Interaction
Experimental	IEP	Inferred from Expression Pattern
Computational Analysis	ISS	Inferred from Sequence or Structural Similarity
Computational Analysis	ISO	Inferred from Sequence Orthology
Computational Analysis	ISA	Inferred from Sequence Alignment
Computational Analysis	ISM	Inferred from Sequence Model
Computational Analysis	IGC	Inferred from Genomic Context
Computational Analysis	RCA	inferred from Reviewed Computational Analysis
Author Statement	TAS	Traceable Author Statement
Author Statement	NAS	Non-traceable Author Statement
Curator Statement	IC	Inferred by Curator
Curator Statement	ND	No biological Data available
Automatically-Assigned	IEA	Inferred from Electronic Annotation
Obsolete	NR	Not Recorded

4.3 Human, Model Organisms & Livestock Species Annotations

In some species, such as model organisms, the availability of tools is generally more complete and direct. On the other hand, for non-model species the analyses are more complex because of lack of annotations, direct experimental data and pathway associations. In these situations extensive comparative mapping approaches must be used to exploit the information that has been gathered on model organisms. Consequently, analyses and interpretation of the results on non-model species might be classified as incomplete, or simply not contain enough 'proven' data for scientific

publications because many inferences would be based on electronic annotations only. These would be very difficult to confirm without any further laboratory experiments.

The status of annotation on livestock species is improving very rapidly mainly due to the availability of genome sequencing technologies. Currently, chicken and cow have been fully sequenced and added to the Ensembl genome browser while the pig genome is currently being sequenced (http://www.ensembl.org). Table 4.2 shows the current statistics on genome annotations of several species; while Table 4.3 represents the assigned species-specific gene ontology annotations. A remarkable aspect of the gene ontology associations is that in the case of cow and chicken, most of GO (>90%) are inferred by electronic annotations (IEA) only.

Table 4.2 Genome Annotation.

(Ensembl gene-builds as of 26 January 2009)

	Human	Mouse	Chicken	Cow	Horse
Assembly	NCBI 36, Oct 2005	NCBI m37, Apr 2007	WASHUC2, May 2006	Btau_4.0, Oct 2007	Equ Cab 2, Sep 2007
Database version	52.36n	52.37e	52.2j	52.4b	52.2b
Base Pairs	3.25E+09	3.42E+09	1.05E+09	3.25E+09	2.43E+09
Known protein-coding genes	21,388	23,019	4,676	20,471	723
Projected protein-coding genes	28	98	6,666	408	14,139
Novel protein-coding genes	9,899	4,918	5,394	175	5,460
Pseudogenes	5,732	3,287	96	686	4,400
RNA genes	388	482	1,026	2,846	1,580
Gene exons	297,252	260,543	182,400	218,492	207,971
Gene transcripts	62,877	48,546	23,316	29,803	28,270
Genscan gene predictions	49,796	49,121	40,505	55,752	107,701
SNPs	15,040,632	14,888,174	2,960,841	2,057,872	

Table 4.3 Gene Ontology Annotations

Statistics as of January 24, 2009 (Appendix 4.1 shows the complete table)

Species	Gene Products Annotated	Annotations	Non IEA Annotations	% of non-IEA Annotations
Arabidopsis thaliana	43,447	113,064	92,940	82.20
Bos Taurus	23,493	100,014	3,758	3.76
Danio rerio	14,910	91,245	22,168	24.30
Escherichia coli	1,597	4,712	4,652	98.73
Gallus gallus	16,334	64,240	2,021	3.15
Homo sapiens	38,846	219,401	64,681	29.48
Mus musculus	17,977	154,280	59,458	38.54
Rattus norvegicus	20,016	161,046	92,810	57.63
Saccharomyces cerevisiae	6,347	85,255	44,416	52.10

Among livestock species, chicken is now recognized as an important model organism for developmental biology and genome research. It shows lots of genetic variation in many different breeds with large population sizes, is easy to breed, and has a high recombination rate (Siegel *et al.* 2006). The importance of chicken as a model organism and its future in research has been reviewed in several papers (Burt 2005; Cogburn *et al.* 2007). Furthermore, the bioinformatics resources for genetic and genome analyses of chicken are developing rapidly. Due to its agricultural importance as a meat and egg production animal, there are vast amounts of data from various sources (i.e. extensive list of quantitative trait studies, development and genomic sequence assembly plus a vast collection of ESTs). The bioinformatics resources for chicken include genome browsers such as ENSEMBL, NCBI, UCSC, and specialized tools for its annotation and gene ontology associations, like AgBase (McCarthy *et al.* 2006). For a detailed

review of the status of chicken resources please refer to Burt and White (2007).

4.4 Methods: Annotation Framework

In order to update and maximize the biological meaning of the experiment, a customized semi-automated re-annotation procedure was performed for all the probes on the array.

4.4.1 Semi-Automated Pipeline Development

Once the statistical analyses of the microarray data were performed, the next step was to develop a framework which would help us to target and associate the probes hybridized on the microarray with gene ontology and pathways. Initially, the sources (were the data was obtained from) were identified and subsequently the procedure of how to integrate multiple sources was set up.

The customized semi-automated re-annotation process integrates various pathways tools and databases, and uses comparative genomics to obtain those pathways and gene associations that might be influenced by the QTL genotypes (Figure 4.1). The framework relates the probes with various gene identifiers, finds gene orthologs in the human genome and links the genes with gene ontologies and pathways (human and chicken pathways). The use of human gene homologies allows the identification of relevant pathways and gene interactions which would be undetected when only using off the shelf tools and poultry specific data.

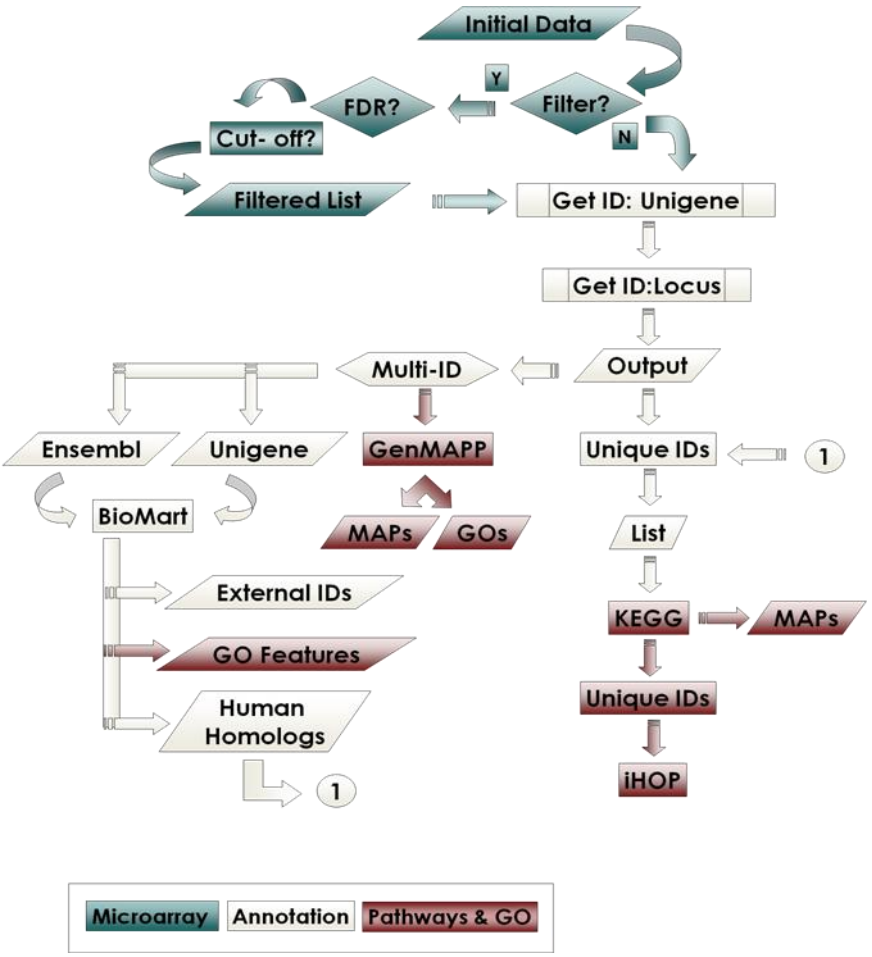


Figure 4.1 Re-annotation framework flow chart

The process assigns gene identifiers, homologies and pathways associations. The framework is mainly divided into three sections. The first section comprises filtering the microarray data according to FDR values and/or related terms annotated to the gene identifiers (*green section*). The second section refers to the semi-automated process of obtaining and integrating the up-dated gene identifiers and their human homologs according to the clone identifiers linked to the microarray probes. Firstly, Get ID: Unigene *Perl* script (Appendix 4.2) obtains the current Unigene annotated to the probes from the NCBI website, the output file created is utilized to obtain the Locus IDs through the process Get ID: Locus *Perl* script (Appendix 4.3). The Get ID: Locus output creates a file with the multiple identifiers. The multiple IDs are then used for the third section of the framework (*red section*) to obtain pathway information and gene ontologies: analysing the dataset with GenMAPP; results are categorized in MAPs (pathways annotated in the GenMAPP chicken database) and GOs (Gene Ontology GenMAPP chicken database). Then the multiple IDs file is used to query BioMart from the Ensembl website using two sets of identifiers (one directly using Ensembl IDs and the second using Unigene IDs). Additionally, from the output generated through Get ID: Locus, a non-redundant gene list (Unique IDs) is created to query KEGG pathways (human and chicken identifiers). Finally, those unique IDs below a 30% FDR are analysed in iHOP to look for direct literature connections.

The genes of the chicken array were annotated by taking the CloneID and GenBank identifiers obtained from the BLAST results; these IDs were used to search for UniGene (Appendix 4.2 Get ID: Unigene) and EntrezGene gene identifiers (Appendix 4.3 Get ID: Locus) directly from the NCBI server. The process was performed using *Perl* scripts. The available gene identifiers were combined into a table containing the multiple identifiers found, facilitating the creation of specific system code gene identifiers lists.

All the genes on the microarray were subsequently queried using BioMart from Ensembl (Hubbard *et al.* 2005) for three purposes: 1) assign features and annotation to the genes and add them to the static annotation, 2) to obtain various alternative gene identifiers, which can be used to run analyses that require a different identifier system code, in order to recognize those genes which could be missed due the annotation differences between databases; and 3) to find gene homologies in humans; genes with an identity greater than 60% were used to run the pathway analyses in order to map genes into model organisms pathways.

4.4.2 Gene Ontologies and Pathway Analysis

Gene ontologies and pathways derived from sources as BioCarta and the Gene Ontology Consortium were documented using GenMAPP (Gene Annotator and Pathway Profiler) (Dahlquist *et al.* 2002). BioMart was used for further gene ontology associations.

The semi-automated pathway analysis uses EntrezGene identifiers to query Kyoto Encyclopedia of Genes and Genomes (KEGG)(Ogata *et al.* 1998) for chicken and human organisms. An important aspect of this analysis is to keep the record and be able to track not only the genes and their gene expression values, but also to identify their genomic location.

Pathway information was recorded for: a) all the genes present per pathway (recording only those pathways where at least one gene of the microarray was present regardless of its expression value); b) the genes documented on those pathways that have been already mapped to the chicken genome; c) the genes identified which were differentially expressed between the two QTL genotypes (Chapter 3) (in this case genes < = 30% FDR value). This was performed by automating the procedure and executing the analysis applying different input criteria (i.e. all genes on microarray, complete list of human homolog genes, significant gene lists) (Appendix 4.4 contains the *Perl* scripts used to format the results obtained from KEGG, the process was divided in two sections, the first section formats the html source code from KEGG results, and the second formats the output file to create a readable table to be integrated into the project database).

Furthermore, for those pathways showing a concentration of significant genes or being functionally related to the trait under study, the locations and functions of the genes and their isoforms that were not present on the microarrays were also recorded (Appendix 4.5 a modified version of the Get ID: Unigene to verify on real time each gene IDs contained per pathway and their genomic location). Furthermore, the differentially expressed genes linked directly to chicken KEGG pathways were searched on iHOP (Information Hyperlinked over Proteins) (Hoffmann & Valencia 2005) to find gene and protein networks derived from the literature.

4.5 Creation of a project-specific Relational Database (RDB)

A relational database (RDB) was created in Access specifically for the targeted genetical genomics approach, containing all the data that was

generated during the development of the project. A relational database organizes the data using common attributes between tables. The database was constantly updated as the analyses were performed.

The initial ('static') annotation, manually curated data and the semi-automated annotations were integrated into the relational database. The RDB also holds the expression profiles of the microarrays, ranging from raw expressions, normalized and statistically analyzed values.

The RDB allows the performance of a two-way analysis. 1) From gene expression towards their functional interpretation (the classical approach), and 2) From 'known' functions and pathways to their gene expression ('reverse' analyses).

4.5.1 Targeted Genetical Genomics Relational Database Design

The targeted genetical genomics experiment database is formed by 17 tables. Figure 4.2 represents the entity relationship diagram of the targeted genetical genomics database, illustrating the relations between tables.

Tables: Description

1. ***Annotation2***: Table retrieved from the original microarray annotation, contains all data describing each of the clones hybridized into the microarray. The data was obtained trough BLAST annotation against various databases (NCBI, ENSEMBL, IPI), includes chromosome locations, ontologies, and human, rat and mouse homologies.

2. ***B_Results_updt***: This information was obtained from the statistical analyses performed in R. Includes the layout of the microarray and the respective statistical values.

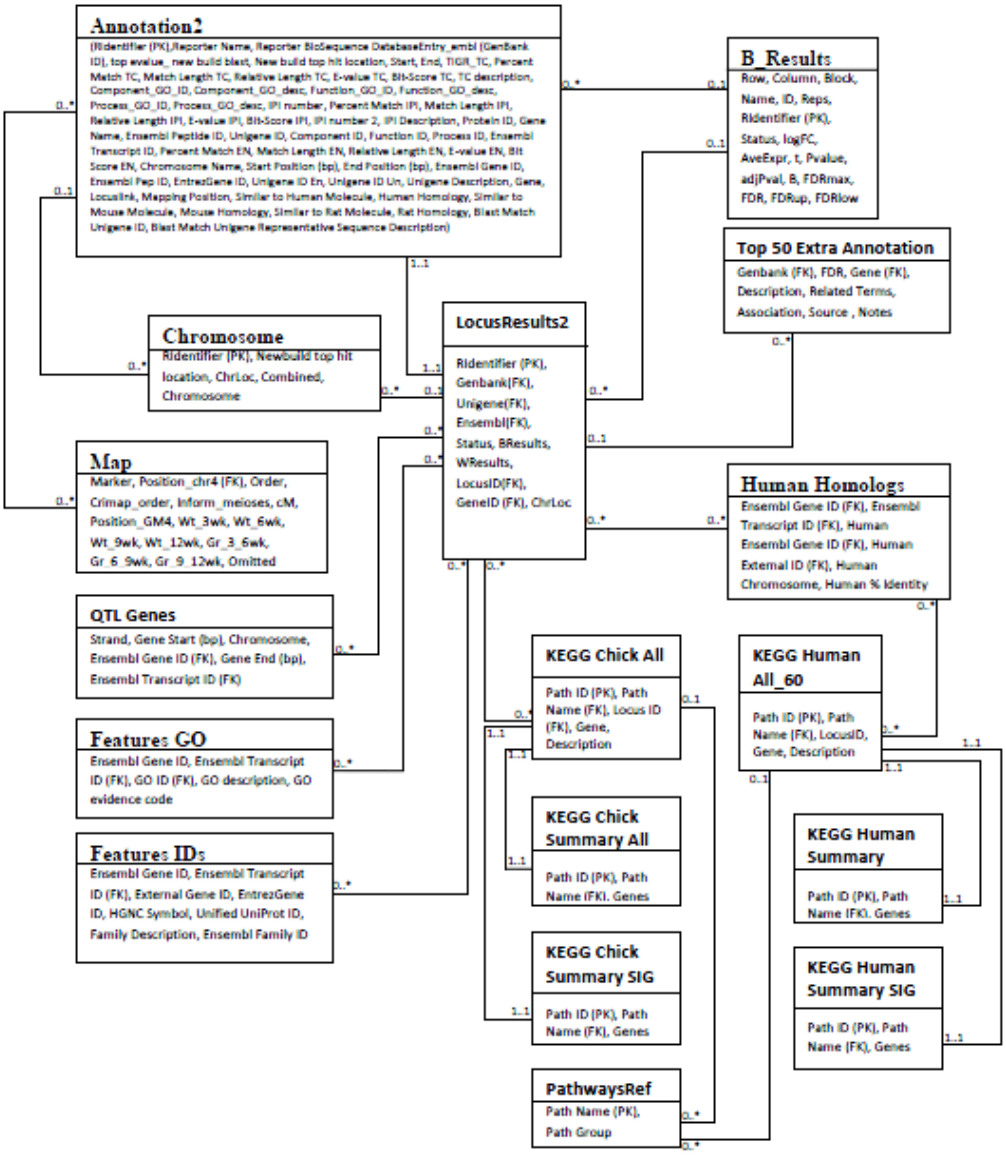
3. ***Chromosome Combined:*** The purpose of this table is to unify, compare and combine the annotations provided for the chromosome locations obtained from various sources. This table combines the original microarray data annotation from the *Annotation2* table and the results found by the semi-automated re-annotation process from the *LocusResults2* table.
4. ***Features GO_Biomart:*** Results from the semi-automated re-annotation framework, concentrates all the gene ontologies for each gene of the microarray. Two sets of gene lists used to query Bioimart gene ontologies; one contains all the genes annotated with an *Ensembl* identifier and the second set includes all the genes annotated with a *Unigene* identifier. The lists of genes were created from the combined results of both annotations (*Annotations2 table* and *LocusResults table*).
5. ***Features IDsChicken:*** Results from the semi-automated re-annotation framework, gathers protein family descriptions and identifiers for which the genes of the microarray encode for, also includes retrieved external identifiers (i.e. LocusLink, EntrezGene ID, HGNC Symbol, Unified UniProt ID) which can be used for further analyses where a specific system-code identifier is required. Follows the same process as the *Features GO_Biomart* table.
6. ***Genes in QTL Biomart:*** Contains all the genes mapped and annotated to the QTL region. The data was obtained through Biomart using the location of the QTL region as delimiters to extract the information. The objective of the table is to obtain all the genes that have been annotated to the region of the QTL, allows targeting of candidate genes through functional annotation without direct gene expression data.

7. ***HomologsHuman_Biomart***: Data retrieved from the re-annotation framework. Holds all the human homologies found for each of the transcripts on the microarrays. Uses the two sets of lists of genes (*Ensembl, Unigene)* to query Biomart homologies, following the same procedure as *Features GO_Biomart* table.
8. ***KEGG Chick All***: Records from the re-annotation framework. Retains all genes from the whole microarray and all the chicken pathways related to them. *Locus* chicken identifiers were used to query KEGG chicken pathways.
9. ***KEGG Chick Summary All:*** Data originates from the *KEGG Chick All table*, the objective of this table is to summarize all the pathways assigned and linked to the genes on the microarray, counting the total number of genes annotated to each of them.
10. ***KEGG Chick Summary SIG:*** Contains only those chicken pathways annotated to the genes considered as differentially expresses at a 30% FDR threshold.
11. ***KEGG Human Summary:*** Data originates from the *KEGG Human all_60 table*. Summarizes human pathways mapped to the differentially expressed genes of the microarray study.
12. ***KEGG Human Summary SIG:*** Data originates from the *KEGG Human all_60 table*. Outlines all the human pathways linked to the complete chicken microarray where the homology identity of the genes was greater than 60%.
13. ***KEGG Human all_60***: Results from the re-annotation framework. Contains all the human pathways related to the complete microarray gene list. Human gene identifiers with a homology identity

percentage greater than 60% were used to query KEGG database. Utilizes *HomologsHuman_Biomart* table to create the input gene list.

14. ***LocusResults2:*** First table obtained from the re-annotation framework, and also considered as the central/main table. A three step semi-automated procedure was performed in order to obtain an up to date *Unigene identifier* and *Locus identifier.* Step 1: prepare the input file containing a set of *clone, genbank,* and *ensembl* identifiers from the *Annotation2 table,* with their respective FDR values obtained from *BResults table.* Step 2: Obtain the updated *Unigene* identifier through a *Perl* script (Appendix 4.2). Step 3: Use as an input the output file created by the Step 2 procedure to create a file containing an up-to-date *Locus* gene identifier and *Chromosome* locations (Appendix 4.3).
15. ***Map 4:*** Holds the marker information and their locations from the QTL mapping results.
16. ***PathwaysRef:*** Reference table, which contains a list of the pathways and the 'parental' modules they are annotated to.
17. ***Top 50 Extra Annotation:*** Manually curated annotation table of the top 50 differentially expressed genes from the targeted genetical genomics study.

Figure 4.2 Database entity relationship diagram



4.6 Results and Conclusion

The creation and use of the relational database helped to centralize the information and execute specific queries for the project. Data centralization facilitates the integration and accession of information that originates from different sources.

Throughout the development of this pipeline several difficulties were encountered, especially as a consequence of the number of candidate genes identified by microarray studies and gene annotation differences between databases and organisms. The pipeline was tested several times, and debugged as necessary. As a result, the RDB holds the available pathway and annotation information for each gene on the microarray. This allows querying all pathways and genes that are present on the microarrays, it also allows detection of pathways which concentrate highly variable genes and investigation of genes which are not hybridized on the microarrays.

In total 603 genes from the microarray were directly related to 118 chicken pathways; the number of genes successfully mapped onto pathways increased when using human gene homologies, where 1,043 genes were found to be linked to 194 pathways. Additionally, pathway annotation of the whole microarray can guide and lead to the identification of genes that might be functionally related to the trait under study but are not present on the microarray. There is a great importance in recognizing and keeping record of the genes being analysed in high-throughput methods and functional analyses, as in many cases, several genes or probes on the microarrays might be encoding for the same enzymes, leading to a misinterpretation of the real candidate gene.

The challenges that researchers now face are increasing constantly; not only is there enormous amounts of data, the data is from different experimental sources, which can vary from experimental results to electronically inferred results derived from bioinformatics tools. The variability and annotation differences between databases and gene/protein identifiers can create serious problems in the functional post-analyses when trying to link the information into gene ontologies and pathways. Also, the application of certain 'significance' cut-offs to the expression values in high-throughput data could miss important expression changes and identification of variability in pathway analyses.

Annotation information for livestock species is very limited, and the methods for the recognition of gene homologies in different species are still under development. Although gene ontology annotations and tools for their analysis are developing rapidly, it is important to remember that gene ontology is the terminology applied independently and accordingly to each investigator's way of expression. They represent terms which can guide the functionality of genes but should not be used as the functional definition of the genes. Furthermore, is necessary to bear in mind that most (>90%) of the GO annotations in livestock are IEA. Therefore drawing conclusions from gene ontologies should be made only to support the hypotheses with other sources.

The utilization of these methods and the utilization of gene homologies can provide an idea of what is expected to be, or be like, but from this to confirmation in the species under study is not a straight forward comparison. The validation on its own species would have to be sustained by diverse literature sources, experimental, and statistical validations. The 'huge' amount of data produced in the post-genomic era is now widely recognised.

It is important to find ways of exploiting the data to its maximum. The developed workflow applied to the targeted genetical genomics study analyses the high-throughput data in a systematic way, integrating gene expressions, pathways, QTLs, and functional annotations utilizing various system code identifiers.

Appendix 4.1 Gene Ontology Statistics

(Statistics as of January 24, 2009)

Species	Gene Products Annotated	Annotations	Non IEA	% non-IEA
Anaplasma phagocytophilum HZ	1,290	3,480	3,480	100.00
*Agrobacterium tumefaciens*str. C58	83	250	250	100.00
Arabidopsis thaliana	43,447	113,064	92,940	82.20
Bacillus anthracis Ames	5,282	13,120	13,120	100.00
Bos Taurus	23,493	100,014	3,758	3.76
Carboxydothermus hydrogenoformans Z-2901	2,611	6,402	6,402	100.00
Caenorhabditis elegans	17,894	94,267	45,831	48.62
Campylobacter jejuni RM1221	1,830	4,658	4,658	100.00
Candida albicans	3,951	18,614	6,468	34.75
Clostridium perfringens ATCC13124	2,892	7,465	7,465	100.00
Colwellia psychrerythraea 34H	4,752	12,126	12,126	100.00
Coxiella burnetii RSA 493	2,033	5,175	5,175	100.00
Danio rerio	14,910	91,245	22,168	24.30
Dehalococcoides ethenogenes 195	1,584	3,958	3,958	100.00
Dictyostelium discoideum	7,238	30,128	19,253	63.90
Drosophila melanogaster	12,537	71,245	55,134	77.39
Escherichia coli	1,597	4,712	4,652	98.73
Ehrlichia chaffeensis Arkansas	1,092	2,868	2,868	100.00
Gallus gallus	16,334	64,240	2,021	3.15
Geobacter sulfurreducens PCA	3,410	8,857	8,857	100.00
Homo sapiens	38,846	219,401	64,681	29.48
Hyphomonas neptunium ATCC 15444	3,109	7,829	7,829	100.00
Leishmania major	3,573	11,441	28	0.24
Listeria monocytogenes 4b F2365	2,819	7,027	7,027	100.00
Magnaporthe grisea	12,876	51,542	29,272	56.79
Methylococcus capsulatus Bath	2,920	7,045	7,045	100.00
Mus musculus	17,977	154,280	59,458	38.54
Neorickettsia sennetsu Miyayama	929	2,439	2,439	100.00
Oomycetes	30	126	126	100.00
Oryza sativa	52,082	64,070	64,070	100.00
Protein Data Bank [multispecies]	20,853	116,158	0	0.00
Plasmodium falciparum	2,208	4,654	4,654	100.00
Pseudomonas aeruginosa PAO1	1,519	7,350	7,350	100.00
Pseudomonas fluorescens Pf-5	3,691	9,711	9,711	100.00

Pseudomonas syringae DC3000	4,006	10,268	10,264	99.96
Pseudomonas syringae pv. phaseolicola 1448[a]	3,506	9,036	9,036	100.00
Rattus norvegicus	20,016	161,046	92,810	57.63
Reactome [multispecies]	257	6,467	6,467	100.00
Saccharomyces cerevisiae	6347	85,255	44,416	52.10
Schizosaccharomyces pombe	5,267	34,491	29,796	86.39
Shewanella oneidensis MR-1	4,843	13,602	13,602	100.00
Silicibacter pomeroyi DSS-3	4,253	10,869	10,869	100.00
Solanaceae	38	68	68	100.00
Trypanosoma brucei	2,978	10,520	10,520	100.00
UniProt [multispecies]	4,451,263	31,289,684	23,059	0.07
Vibrio cholera	3,858	9,430	9,430	100.00

Appendix 4.2 Get ID: Unigene

```
#! /usr/bin/perl -w
use strict;
use warnings;
require  LWP::UserAgent;
my $ua = LWP::UserAgent->new;
   $ua->env_proxy();

# Claudia Cabrera
# The program takes in a automated way the unigene IDS of the genes
# providing a clone ID or genbank ID
# The input list must be tab delimited and without header row and
# must contain
# 1st column : clone id
# 2nd column : geneBank ID
# 3rd column : ENSEMBL ID
# 4th column : Status, were C means that its value its high in the
within analysis
# 5th column : Between Results FDR
# 6th column : Within Results FDR

############## Start: Input Example ####################
#ChEST905J10   BU225400       ---                   C      0.531630695    7.18E-03
#ChEST774I17   BU389307       ENSGALG00000004780    C      0.447075512    7.34E-03
#ChEST879L18   BU285868       ---                   C      0.430180209    7.65E-03
#ChEST805A5    BU271238       ---                   C      0.395511358    7.75E-03
############### End: Input Example ######################

############## Start: Output Example ###################
#RIdentifier   Genbank        UniGene        Ensembl        Status  BResults
       WResults
#ChEST905J10   BU225400       NULL           ---            C       0.531630695
       7.18E-03
#ChEST774I17   BU389307       Gga.21386      ENSGALG00000004780     C
       0.447075512    7.34E-03
#ChEST879L18   BU285868       Gga.40082      ---            C       0.430180209
       7.65E-03
#ChEST805A5    BU271238       Gga.21059      ---            C       0.395511358
       7.75E-03
############### End: Output Example ###################

# Takes the input file as an argument in the command line
my $filename1=$ARGV[0];

#open input file
open(FILE1, $filename1) or die "Cannot open file
\"$filename1\"\n\n";

#Declares output file
my $out="UnigeneIDs.txt";

#Creates a file to make the output
open(FO,">$out") or die "Cannot create file \"$out\"\n\n";

# Insert headers into the output file
print FO
"RIdentifier\tGenbank\tUniGene\tEnsembl\tStatus\tBResults\tWResults\
n";

#Introduces the content to an array
```

```
my @list=<FILE1>;
my $cont=0;
foreach my $id(@list){
	$cont=$cont+1;
	chomp $id;
	$id =~ s/\r$//;
	$id =~ s/^\s+//;
	$id=~ s/\s+$//;
	my @two= split('\t',$id);
	my $clone=$two[0];
	my $gb=$two[1];
	my $en=$two[2];
	my $status=$two[3];
	my $Bresult=$two[4];
	my $Wresult=$two[5];
	# Sends the internet content request
	my $request = HTTP::Request-
>new('Get',"http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene&cmd=
search&term=$clone");
	#my $request = HTTP::Request-
>new('Get',"http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homolog
ene&cmd=search&term=ChEST370A4");

	# Gets Internet response content
	my $response = $ua->request($request);
	my $seq=$response->content;

	# Formats the response content
	if ($seq =~ m/id\=\"Gga\..*\"\>(.+)\<\/a\>/)
	{
		my $ugid = $1;
		print FO $clone,"\t",$gb,"\t",$ugid,"\t",$en,
			"\t",$status,"\t",$Bresult,"\t",$Wresult,"\n";
		print "$cont $clone: $ugid**\n";
	}
	else {
		my $request = HTTP::Request-
>new('Get',"http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene&cmd=
search&term=$gb");
		my $response = $ua->request($request);
		my $seq=$response->content;
			if ($seq =~ m/id\=\"Gga\..*\"\>(.+)\<\/a\>/)
			{
				my $ugid = $1;
				print FO
$clone,"\t",$gb,"\t",$ugid,"\t",$en,

	"\t",$status,"\t",$Bresult,"\t",$Wresult,"\n";
				print "$cont $gb: $ugid**\n";
			}
			else {
				my $ugid = "NULL";
				print FO
$clone,"\t",$gb,"\t",$ugid,"\t",$en,
	"\t",$status,"\t",$Bresult,"\t",$Wresult,"\n";
				print "$cont $gb: $ugid**\n";
			}
	}
}
```

Appendix 4.3 Get ID Locus

```
#! /usr/bin/perl -w
use strict;
use warnings;
require  LWP::UserAgent;
my $ua = LWP::UserAgent->new;
   $ua->env_proxy();

Claudia Cabrera
# The program takes in a automated way the locus IDS of the genes
# providing as an input the results of the program getunigeneid.pl
# or the following list:
# The input list must be tab delimited and without header row and
# must contain
# 1st column : clone id
# 2nd column : geneBank ID
# 3rd column : Unigene ID
# 4th column : ENSEMBL ID
# 5th column : Status, where C means that its value its high in the
within analysis
# 6th column : Between Results FDR
# 7th column : Within Results FDR

############## Start: Input Example ####################
#ChEST291J16   BU224351       Gga.10  ---     OK      0.680912469    0.185713228
#ChEST912I10   BU207821       Gga.100         ENSGALG00000006351     OK
       0.434481545    0.131481693
#ChEST997F19   BU212557       Gga.100 ---     OK      0.702268274    0.35235636
#ChEST51H23    BU218004       Gga.10003       ---     OK      0.313105437
       0.360032876
############## End: Input Example ######################

############## Start: Output Example ###################
#RIdentifier   Genbank UniGene        Ensembl               Status  BResults
       WResults       LocusID GeneID  ChrLoc
#ChEST291J16   BU224351       Gga.10          ---                                     OK
       0.680912469    0.185713228    395191  OTX2    5
#ChEST912I10           BU207821       Gga.100         ENSGALG00000006351     OK
       0.434481545    0.131481693    395274  FAS             6
#ChEST997F19           BU212557       Gga.100         ---
       OK      0.702268274    0.35235636     395274  FAS             6
#ChEST51H23                   BU218004       Gga.10003       ---
              OK      0.313105437    0.360032876    416421  MKL2    14
############## End: Output Example ###################

# Reads input file
my $filename1=$ARGV[0];

#open file
open(FILE1, $filename1) or die "Cannot open file
\"$filename1\"\n\n";

#Declares output file
my $out="LocusResults.txt";

#creates a file to make the output
open(FO,">$out") or die "Cannot create file \"$out\"\n\n";

print FO
"RIdentifier\tGenbank\tUniGene\tEnsembl\tStatus\tBResults\tWResults\
tLocusID\tGeneID\tChrLoc\n";
```

```
#Introduces the content to an array
my @list=<FILE1>;
my $cont=0;
foreach my $id(@list){
      $cont=$cont+1;
      chomp $id;
      $id =~ s/\r$//;
      $id =~ s/^\s+//;
      $id=~ s/\s+$//;
      my @two= split('\t',$id);
      my $clone=$two[0];
      my $gb=$two[1];
      my $ugid=$two[2];
      my $en=$two[3];
      my $status=$two[4];
      my $Bresult=$two[5];
      my $Wresult=$two[6];
      my $gene="NULL";
      my $chr="NULL";
      my $locus="NULL";
      if ($ugid ne "NULL"){
            my $request = HTTP::Request-
>new('Get',"http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=sum
mary&TermToSearch=$ugid");
            my $response = $ua->request($request);
            my $seq=$response->content;
            if ($seq =~
m/\<b\>\d+\:\s\<\/b\>\<\/td\>\<td.*\"\>(.+)\<\/a\>/)
            {
                  $gene = $1;
            }
            if ($seq =~ m/Chromosome\: \<\/strong\>(\d+|\w+)/)
            {
                $chr = $1;
            }
            if ($seq =~ m/GeneID\: \<\/strong\>(\d+)\<\/td\>/)
            {
                  $locus= $1;
            }

            print FO $clone,"\t",$gb,"\t",$ugid,"\t",$en,
            "\t",$status,"\t",$Bresult,"\t",$Wresult,"\t",
                              $locus,"\t",$gene,"\t",$chr,"\n";

            print "$cont:$clone\t$locus\t$gene\tChr $chr\n";
            }

      else {
            my $locus = "NULL";
            my $gene = "NULL";
            my $chr = "NULL";
            print FO $clone,"\t",$gb,"\t",$ugid,"\t",$en,
                  "\t",$status,"\t",$Bresult,"\t",$Wresult,"\t",
                  $locus,"\t",$gene,"\t",$chr,"\n";
            print "$cont: $clone\t$locus\t$gene\tChr $chr\n";

      }
}
```

Appendix 4.4 KEGG Format

```
#! /usr/bin/perl -w
use strict;
use warnings;

# Claudia Cabrera
# Format KEGG results part 1
# Takes as an input file the source code results from KEGG

################### Start: Input File Example ######################
#
#<html><head><title>
#Pathway Search Result
#</title>
#<link rel="stylesheet" href="/css/kegg.css" type="text/css">
#</head>
#<body>
#<h2>
#Pathway Search Result
#...
#<hr>
#<li><a href="/kegg-bin/mark_pathway_www?23004/gga03010.args"><b>gg…
#396262 RPLP1; ribosomal protein, large, P1 [SP:RLA1_CHICK]
#415551 RPL4; ribosomal protein L4 [SP:Q5ZII1_CHICK]
#419895 RPL10A; ribosomal protein L10a [SP:Q6EE62_CHICK]
#419904 RPS10, LOC419904; ribosomal protein S10
#427675 RPS15A, LOC427675; ribosomal protein S15a
#<li><a href="/kegg-bin/mark_pathway_www?23004/gga00010.args"><b>gga00010 Glycolysis
/ Gluconeogenesis</b></a>
#374193 GAPDH; glyceraldehyde-3-phosphate dehydrogenase [EC:1.2.1.12]
################### End: Input File Example ########################

################### Start: Output File Example ######################
#PATH ID       PATH Name      LocusID Gene    Description
#gga00230      Purine metabolism      396540  ADSL    adenylosuccinate lyase
[EC:4.3.2.2] #gga00230 Purine metabolism      415635  POLR2C polymerase (RNA) II (DNA
directed) #gga00230    Purine metabolism      415732  ADCY7, LOC415732       adenylate
cyclase 7 #gga00230    Purine metabolism      416058  RCJMB04_1j11, IMPDH2   IMP
(inosine
#gga00230      Purine metabolism      416710  POLR1B, LOC416710      polymerase (RNA) I
################### End: Output File Example ########################

# Takes as an argument in the command line a file containing the
html results of KEGG
my $filename1=$ARGV[0];
my ($ptid,$pname);

#open file
open(FILE1, $filename1) or die "Cannot open file
\"$filename1\"\n\n";

#Declare output file
my $out="PATHsHUMAN_ALL_60.txt";

#creates a file to make the output
open(FO,">$out") or die "Cannot create file \"$out\"\n\n";

# Insert headers into the output file
print FO "PATH ID\tPATH Name\tLocusID\tGene\tDescription\n";

#Introduces the content to an array
my @list=<FILE1>;
my $cont=0;
```

```
foreach my $id(@list){
      $cont=$cont+1;
            if ($id =~
m/\<li\>\<a.*\"\>\<b\>(\w+\d+)\s(.+)\<\/b\>\<\/a\>/)
            {
                  $ptid = $1;
                  $pname = $2;
            }

            if ($id =~ m/(\d+)\s(.+)\;\s(.+)\n/)
            {
            my $loc = $1;
            my $gene = $2;
            my $des = $3;
            print FO
$ptid,"\t",$pname,"\t",$loc,"\t",$gene,"\t",$des,"\n";
            #print $cont,":",$ptid,"\t",$pname,"\n";
            }

 }
```

PART 2

```
#! /usr/bin/perl -w
use strict;
use warnings;

# Claudia Cabrera
# Format KEGG results part 2
# takes gene column and splits it into each of them

################### Start: Input File Example ######################
#PATH ID       PATH Name      LocusID Gene    Description
#gga00230      Purine metabolism      396540  ADSL    adenylosuccinate lyase
[EC:4.3.2.2] #gga00230 Purine metabolism      415732  ADCY7, LOC415732       adenylate
cyclase 7 #gga00230    Purine metabolism      416058  RCJMB04_1j11, IMPDH2   IMP
(inosine
#gga00230      Purine metabolism      416710  POLR1B, LOC416710      polymerase (RNA) I
################### End: Input File Example ########################

################### Start: Output File Example ######################
#gga00230      Purine metabolism      396540  ADSL    adenylosuccinate lyase
[EC:4.3.2.2] #gga00230 Purine metabolism      415732  ADCY7   adenylate cyclase 7
[EC:4.6.1.1]
#gga00230      Purine metabolism      415732  LOC415732      adenylate cyclase 7
[EC:4.6.1.1]
#gga00230      Purine metabolism      416058  RCJMB04_1j11   IMP (inosine
monophosphate) #gga00230      Purine metabolism      416058  IMPDH2  IMP (inosine
monophosphate) dehydrogenase #gga00230       Purine metabolism      416710  POLR1B
       "polymerase (RNA) I polypeptide B, 128kDa #gga00230  Purine metabolism
       416710  LOC416710      "polymerase (RNA) I polypeptide B, ###################
End: Output File Example ########################

my $filename1=$ARGV[0];
#open file
open(FILE1, $filename1) or die "Cannot open file
\"$filename1\"\n\n";

#output file
my $out="PATHsHUMAN_ALL_60part2.txt";

#creates a file to make the output
open(FO,">$out") or die "Cannot create file \"$out\"\n\n";
```

```
#Introduces the content to an array
my @list=<FILE1>;
my $cont=0;
foreach my $id(@list){
      $cont=$cont+1;
      my @two= split('\t',$id);
      my $pathID=$two[0];
      my $pathName=$two[1];
      my $locID=$two[2];
      my $gene=$two[3];
      my $des=$two[4];

      my @gp= split(",",$gene);
      foreach my $gs(@gp){
            $gs =~ s/^\s+//;
            print FO
$pathID,"\t",$pathName,"\t",$locID,"\t",$gs,"\t",$des;
            print $gs,"\n";
      }
}
```

Appendix 4.5 Pathway ID

```
#! /usr/bin/perl -w
use strict;
use warnings;
require  LWP::UserAgent;
my $ua = LWP::UserAgent->new;
   $ua->env_proxy();

# Claudia Cabrera
# The program takes in a automated way the locus IDS of the genes
# providing as an input the results of the program getunigeneid.pl
# or the following list:
# The input list must be tab delimited and without header row and
# must contain 1st column : locus id
####Program used to look for the annotated genomic locations of all
genes and isoforms per pathway

########### Start: Input file example : All Locus Ids annotated to 1 Pathway #######
#418312
#418968
#396286
#408026
#418561
#420681
#420947
########### End: Input file example ##########

########### Start: Output file example #######
#Locus ID      Chromosome
#418312        1
#418968        1
#396286        1
#408026        1
#418561        1
########### End: Output file example #########
```

```
# Reads input file
my $filename1=$ARGV[0];

#open file
open(FILE1, $filename1) or
        die "Cannot open file \"$filename1\"\n\n";

my $out="TyrosineRES.txt";

#creates a file to make the output
open(FO,">$out") or die "Cannot create file \"$out\"\n\n";

print FO "LocusID","\t","ChrLoc\n";

#Introduces the content to an array
my @list=<FILE1>;
my $cont=0;
foreach my $id(@list){
      $cont=$cont+1;
      chomp $id;
      $id =~ s/\r$//;
      $id =~ s/^\s+//;
      $id=~ s/\s+$//;
      my $ugid=$id;
      my $request = HTTP::Request-
>new('Get',"http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=sea
rch&term=$ugid");
      my $response = $ua->request($request);
      my $seq=$response->content;
      $seq =~ m/\<strong\>chromosome\<\/strong\>(.*)\<a/;
      my $chr = $1;
      print FO $ugid,"\t",$chr,"\n";
      print "$cont:$ugid\n";
}
```

Chapter 5 Post-Analyses

5.1 Introduction

In this chapter the focus is to investigate *how* high-throughput data can be further analysed and interpreted by performing post-analyses. The objective of the post-analyses is to identify and characterize further, through bioinformatics procedures, the functionality of the 'interesting' gene lists and/or genetic regions targeted by high-throughput technologies.

The methodology presented is based on the integration of technologies. In this chapter I describe how to perform an in-depth analysis of a genetic region of interest and how to detect global genetic variation. The in-depth analysis allows better characterization of the candidate genes which are very poorly annotated and it also permits identification of previously reported QTLs on the syntenic regions of various species (i.e. chicken, cow, mouse, rat, human). Whilst, the global analysis allows to detect the downstream effects (flow of causality) through functional variation found in pathways.

To exemplify this, the post-analyses were applied to the gene expression profiles and annotations of the chicken body weight QTL study (presented in Chapter 3 and further annotated in Chapter 4). The targeted genetical genomics study investigated the differences within a population of an advanced inter-cross between broiler and layer chickens with a known QTL responsible for body weight. The study allowed the identification of potential positional candidate genes that might be responsible for chicken body weight. However, the positional candidate genes annotation was not informative and in order to investigate further their relationship to the trait under study more focused analyses are required. In addition, the QTL region

of the targeted study was subject of the focused genomic region analyses (integration of physical and linkage maps and syntenic regions).

The applied methods considered in this study involve the data generated by various techniques (BLAST, filtering, annotation integrations, sequence analyses, gene ontologies and pathway analysis) performed via public domain sources used commonly in genome research, such as NCBI, ENSEMBL, UCSC, KEGG, GenMAPP, Genious, Animal QTLdb, and map viewers (bioinformatics tools and methods are explained in Chapter 2).

5.2 Methods

The integration of data from several public domain resources and the amount of information gathered by the previous studies requires a systematic approach to be able to make the data more useful. In this case an analysis-framework was developed (Figure 5.1). The analysis-framework consists on the methods followed to integrate the probe microarray annotations to external genetic resources. Briefly, after the genes were further annotated, the gene sequences and homology gene identifiers were utilized to perform comparative genomics and to identify synteny regions in other species (human, cow, rat, and mouse). This allowed the identification of previously reported QTLs in the synteny regions. The gene homology identifiers were used to target pathways annotated in humans. The post-analyses performed fall into two categories: 1) an in-depth analysis of the QTL genomic region; and 2) a global analysis.

In addition, the functional analyses were performed in two ways. 1) The 'classic' approach: going from the gene expression towards their functional

interpretation, and 2) the 'reverse' approach: going from known function to the analysis of gene expression differences.

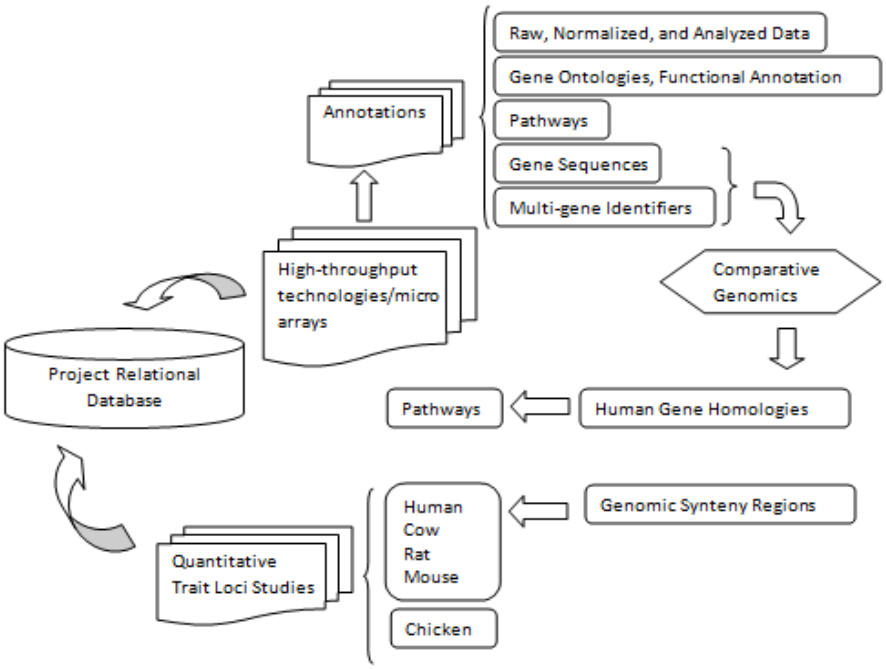


Figure 5.1 Integrating technologies in genome research workflow.

Overview of the integrated processes: 1)further microarray annotations (GO, pathways, genetic sequences, and various IDs); 2) Use these sequences and IDs to perform comparative genomics and obtain human gene homologies and synteny regions for human, rat, cow, mouse in order to look for QTLs in those regions across species

5.2.1 In-depth QTL genomic region analyses

In the following sections in-depth QTL genomic region analyses are presented. These include the investigation on genomic sequences performing sequence analyses, BLAST techniques, integration of physical and linkage maps and also the investigation of the QTL syntenic regions in other species.

5.2.1.1 Sequence analyses

In the previous study, four genes were considered as potential positional candidates, because they mapped to the QTL region under investigation and also were considered as significant differentially expressed between the alternative genotypes of the QTL (Chapter 3). A detailed analysis of their putative functions was carried out. The conservation across species of the candidate genes was observed and their protein domains, functional sites and alternative splicing were investigated.

Initially, the nucleotide sequences of the four candidate genes were extracted from NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuccore). The sequences were investigated using BLAST in NCBI and ENSEMBL to obtain sequences with similar regions, their putative functions, and annotations (Appendix 5.1). The next step was to query the targeted genetical genomics database for other clones hybridized on the microarray that might be encoding for the same candidate genes and investigate their expression profiles. This was performed by looking at the genomic region of the candidate genes at the UCSC genome browser (http://genome.ucsc.edu/). The identifiers of the ESTs located in the same region as candidate genes were extracted to perform the query. In addition, if the probe on the microarray was an un-spliced EST the nearest ESTs to this region were also observed, by obtaining the other clone identifiers which were also queried against the targeted genetical genomics database.

The conservation across species was observed also using the UCSC browser in order to identify genetic regions that might contain functional elements. If the conservation across species showed variability on the coding regions of the genomic sequence, an '*in-silico*' protein modification using Geneious (© 2005 -2008 Biomatters Ltd.) was performed to test if the variability or splice

variants alter the gene 'functionality' or resulting conserved domains of the sequence.

The *in-silico* modification consisted of extracting the genomic sequence of the region of interest and recording the coordinates of each exon. The sequences were modified manually. Various models (scenarios) were created utilizing different exon combinations. The models were created according to the exon sequence conservation and splice variants. We can take as an example two ESTs (EST*x* and EST*y*) that map to the same genomic location. EST*x* contains four exons (*a*, *b*, *c*, and *d*), and EST*y* contains only three exons (*a*, *c*, and *d*). In addition, exon *c* contains splice variants (less nucleotides in EST*x* than in EST*y*) (Figure 5.2). If the exon is always present and conserved on the various ESTs of the same genomic region, this exon would not be omitted on the models of the *in-silico* modifications (e.g. exon *a* and *d* are constant, therefore they are always included in the sequence models with no modifications). But, if the exon presents sequence splicing variants (e.g. exon *c*) and/or alternative presence (e.g. exon *b*) this would drive to the creation of 'sequence models'. One model '*Model 1'* would be the same sequence as EST*x* but omitting exon *b,* as in EST*y.* A second model '*Model 2'* would be based in the sequence of EST*y* although with the alternative splice variant of EST*x* for exon *c* (Figure 5.2). Each of the resulting *model* sequences was translated into an amino acid sequence on a six-frame reading frame option. The resulting amino acid sequences were BLAST against the NCBI conserved domains database (Marchler-Bauer *et al.* 2007) to investigate if the sequences could conserve functionality after modification.

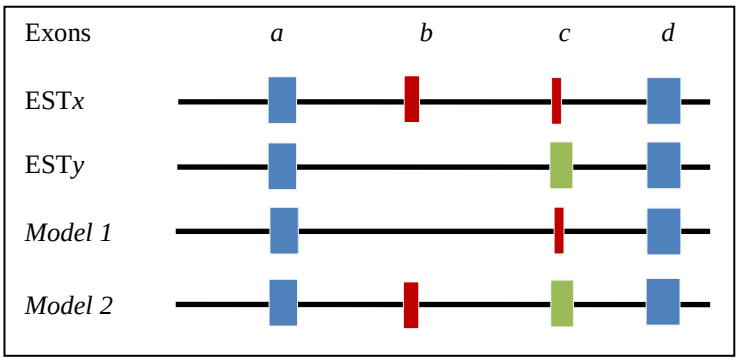


Figure 5.2 *In-silico* DNA sequence modification.

Two ESTs (EST*x* and EST*y*) map to the same genomic location. EST*x* contains four exons (*a*, *b*, *c*, and *d*), and EST*y* contains only three exons (*a*, *c*, and *d*). In addition, exon *c* contains splicing variants (less nucleotides in EST*x* than in EST*y*). The sequence alternatives drive to the creation of nucleotide sequence models. *Model 1*: same sequences as EST*x*, but no exon *b*. *Model 2:* same sequence as EST*y* but with exon *c* variant from EST*x*.

5.2.1.2 Integrating Physical and Linkage maps

The integration of the physical and linkage maps allows to compare the QTL under study with other chicken QTLs found on the same region. The genomic locations of the flanking markers used on the targeted study to map the body weight QTL were obtained utilizing the sequence and the SNP identifiers to query the single nucleotide polymorphism NCBI database (http://www.ncbi.nlm.nih.gov/SNP/). The annotated coordinates of the SNPs used as flanking markers on the study are based on the physical map and therefore they are annotated in base pairs (bp). The physical map information was viewed and obtained through the NCBI MapViewer tool (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9031). The chicken linkage map data were obtained from the ChickenQTLdb (http://www.animalgenome.org/QTLdb/chicken.html) (Hu & Reecy 2007). This database is based on Groenen *et al.,* (2000) chicken consensus linkage map and refers to its coordinates on centiMorgans (cM).

Subsequently, based on the most proximate common markers between the physical sequence map and the chicken linkage consensus map, the flanking markers' coordinates were obtained in centiMorgans. The physical map was used as a base for the coordinates of the QTL. The order of the markers between the physical map and the linkage map was compared.

Once the location was transformed to cM on the consensus map it was possible to query the chicken QTL database (Hu & Reecy 2007). The database was queried according to the region of the body weight QTL under study and the co-localized QTLs mapped in other studies were obtained. Additionally, the regions and gene expression of the markers within the interval region of the QTL on the linkage map, but mapping outside the QTL region on the physical map were also investigated.

5.2.1.3 Synteny Regions

The human synteny sections of the targeted QTL region under study were documented through Ensembl. Once recognizing the synteny sections in the human genome the data were linked to the synteny regions of other organisms (mouse, rat and cow). The QTLs that have been reported on the other organisms synteny regions and the associations were investigated through the rat genome database (RGD; http://rgd.mcw.edu/). The RGD database is a collaborative project with the goal to make publicly available rat genetic and genomic research efforts. RGD integrates functional gene annotations, mapping, disease and phenotypic data (i.e. QTL reports, markers, ESTs, pathway and ontologies). (Twigger *et al.* 2007).

5.2.2 Global Analysis

The global analysis focuses on the downstream effects of the QTL. The functional role of all the significant genes regardless of their genomic

position was investigated allowing the observation of changes that the different QTL genotypes cause in other genes. A functional global analysis can identify if the genetic variation is concentrated on particular pathways, gene functions, processes, or cellular components.

In this chapter the significant genes refer to the 580 differentially expressed transcripts from the targeted genetical genomics study (Chapter3); however, only 511 transcripts could be associated with some kind of annotation, and 368 were linked to an EntrezGene identifier. Chromosome 4 annotations contain approximately 1,187 genes of which 498 genes (40 differentially expressed) where represented on the microarray [unpublished data].

5.2.2.1 Pathways and Gene Ontologies

The pathways that were enriched for the differentially expressed genes were grouped into their main categories (or major pathway groups) such as cell communication, translation, amino acid metabolism, and signal transduction. Also, there is a high probability that the expression of some highly relevant genes will not be measured because they are not represented on the microarray. One indirect approach to address this problem was to look within the module categories with higher concentration of differentially expressed genes and investigate the functions of those genes which are not hybridized onto the microarray. Additionally, Ontologizer (Bauer *et al.* 2008) and AgBase (McCarthy *et al.* 2006) were utilized to analyze the overrepresentations of the gene ontologies associations.

5.3 Results and Discussion

5.3.1 In-depth QTL genomic region analyses

The focused analysis of the QTL on GGA4 (23 Mbp ~ 37 Mbp) from the targeted genetical genomics study (Chapter 3) identified four differentially expressed genes on the microarray which had no direct meaningful annotation available (BU452163; BU415609; BU465968; and BU463895).

5.3.1.1 Sequence Analysis

Only three differentially expressed genes were functionally identified under the QTL: (1) BU452163, *inturned planar cell polarity effector homolog* (INTU); (2) BU415609, clone with 76% identity with the human PHF17 also known as JADE1; (3) BU465968 (LOC428728), gene with an identity of 68% to human *alpha-aminoadipate aminotransferase* (AADAT). For BU463895 no meaningful annotation could be obtained, although it is located in the upstream genomic region of *zinc finger protein 827* (ZNF827), and has a 100% identity overlap with 59 base pairs of BU385281 coding for ZNF827 (Appendix 5.1).

The UCSC genomic sequences analysis performed on the positional candidate's transcripts did not find any other co-localized ESTs on the microarray. The sequence analysis performed on the four candidate genes demonstrated conservation of the coding regions across species (Appendix 5.1). However, variability and alternative splicing were observed on various exons of the BU465968 clone whose sequence is located on the AADAT genomic region (Figure 5.3).

The genomic sequence of AADAT is annotated with 13 exons (based on ENSEMBL) and 6 conserved domains were linked to the protein sequence: (1) COG3977,*alanine-alpha-ketoisovalerate aminotransferase* (amino acid

transport and metabolism); (2) ARO8, *transcriptional regulators containing a DNA-binding HTH domain* and an *aminotransferase domain* (MocR family) and their eukaryotic orthologs (transcription/amino acid transport and metabolism); (3) COG0436, *aspartate/tyrosine/aromatic aminotransferase* (amino acid transport and metabolism); (4) HisC, *histidinol-phosphate/aromatic aminotransferase* and *cobyric acid decarboxylase* (amino acid transport and metabolism); (5) Aminotran_1_2, *aminotransferase class I and II*; (6) MalY, *bifunctional PLP-dependent enzyme with beta-cystathionase* and *maltose regulon repressor activities* (amino acid transport and metabolism).

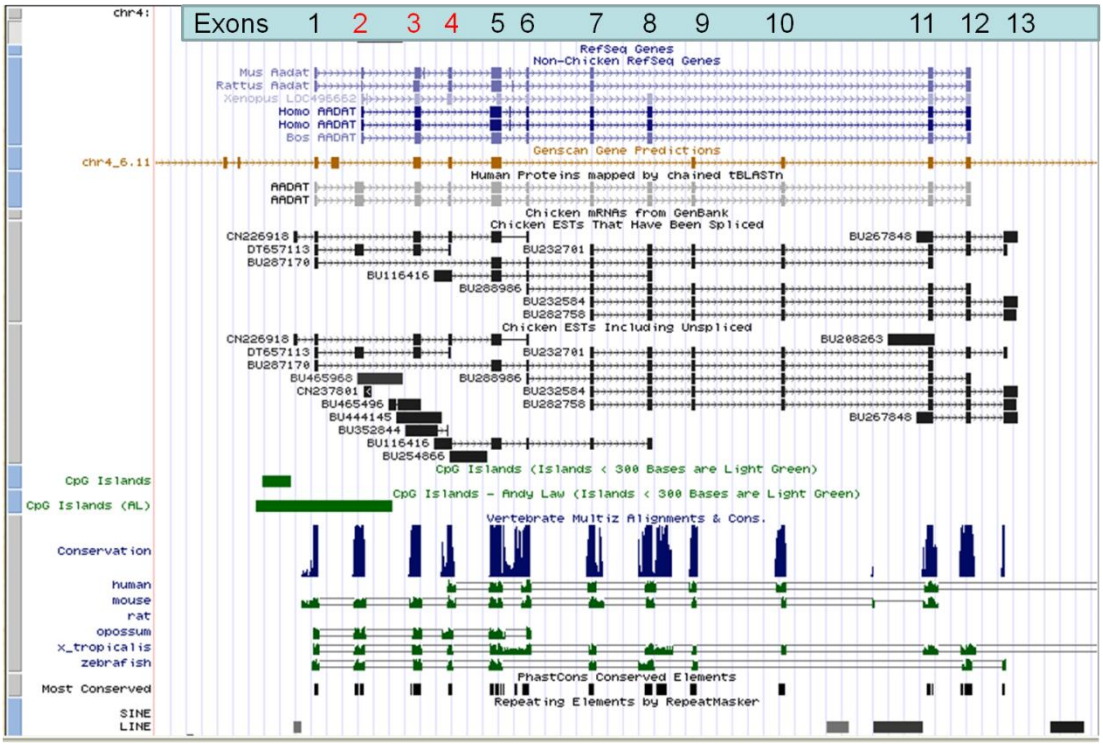


Figure 5.3 AADAT genomic visualization.

UCSC Genome browser visualization (http://genome.ucsc.edu/) of AADAT genomic region, showing alignments across species, alternative splicing and conservation. The first section of the image (blue sequences) illustrates AADAT genomic sequences alignments across different species. The second section (black sequences) displays various chicken ESTs that have been alternatively spliced. The third section shows multiple alignments and their conservation within chicken and across species.

Three exons (exons 2, 3 and 4) from the 13 exons present in the sequence showed variability among the ESTs. Five *in-silico* sequence *models* were created with alternative exon combinations (Figure 5.3). As a result of the *in-silico* modification, the deletion of the sequence of AADAT-exon-four (Figure 5.4) resulted on a protein coding sequence which however presented a disruption on the *aminotransferase class I and II* conserved domain (Figure 5.5). The 'original' sequence has 425 residues, whilst the sequence without exon-four contains 403 residues. A total 25 residues (124 -148) were deleted.

```
Alignment result:

        121                                                        128

Exons    LCKVFEMLINPGDSILLDAPTYSGTLAALRPLGCSIINVPSDQHGIIPKALKEILSAWSP

NoExon4 LCK------------------------LRPLGCSIINVPSDQHGIIPKALKEILSAWSP
```

No further 'protein' coding sequences were found in any of the 30 resulting amino acid sequences obtained from the *in-silico models.*

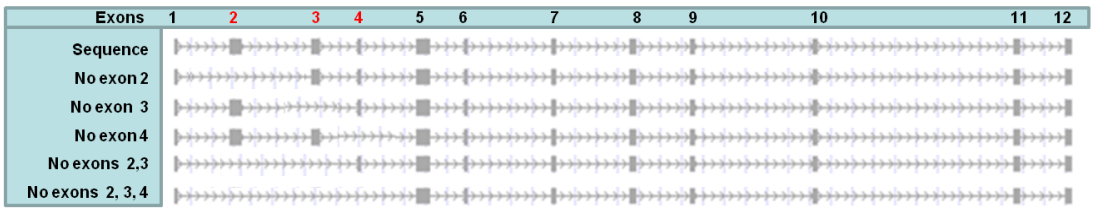


Figure 5.4 *In-silico* models of AADAT sequence.

The sequence contains 13 exons (picture displays only from 1-12). The exons 2 – 4 showed variability in the ESTs sequences, driving to the creation of five models.

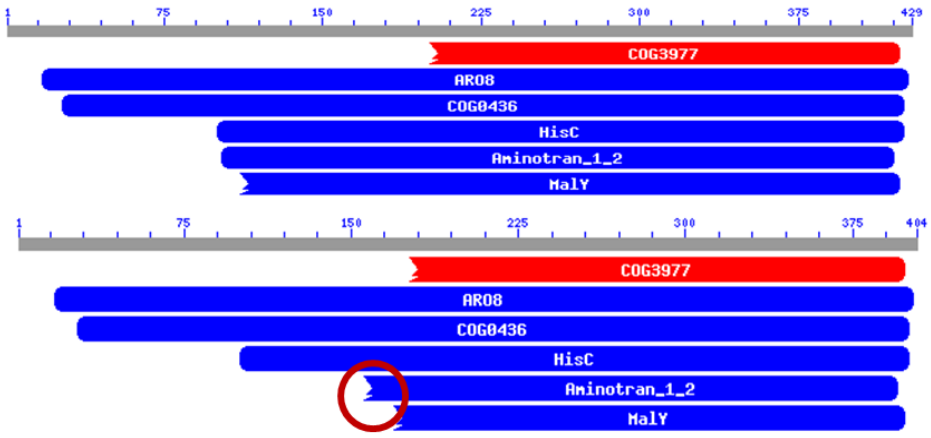


Figure 5.5 Conserved Domains.

The top frame represents the normal protein sequence of AADAT and the recognized conserved domains. The bottom frame represents the disruption of the *aminotransferase class I and II* conserved domain due to the in-silico deletion of exon four.

The biological function of the three characterized positional candidate genes (INTU, PHF17, AADAT) was further investigated. INTU was found to contain a PDZ domain involved in protein binding (Nagase et al. 1999). The PHF17 gene is related to various biological processes including regulation of transcription, apoptosis and negative regulation of cell growth (Zhou et al. 2002). Human AADAT is involved in lysine degradation, lysine biosynthesis and tryptophan metabolism, with ubiquitous expression that is highest in the liver (Goh et al. 2002). It is important to mention that previous experiments demonstrated a strong relation between lysine deficiency and decreased body weight and tissue protein levels. A disruption of the supply of a single essential amino acid can disturb growth mechanisms (Tesseraud et al. 1996). Moreover, AADAT was expressed at higher levels in the *QQ* muscle

genotype involved in lysine biosynthesis which could have a positive impact on the increase of body weight (Fatufe et al. 2004).

5.3.1.2 Integrating Physical and Linkage maps

The body weight QTL region was linked to the consensus linkage map. The integration of linkage and physical maps was complicated by differences in the marker order between different maps.

The most proximal common markers between the maps indicated the position of the QTL on the consensus linkage map to localize around ~ 78cM to 118 cM (Figure 5.6). The lower flanking marker of the body weight QTL (snp.28.110.2096.S.1) is located at the ~23Mbps, and the most proximal marker localized on both maps is LEI0095 (25Mbps; 78cM). The upper flanking marker (snp.3.260.3284.S.2) was mapped to ~ 38 Mbps, being ADL0194 the closest common marker (38Mbps; 118 cM). As expected, differences were found on the order of the markers between the two maps (Figure 5.6). This could be simply due genotyping errors in different experiments, and additionally it has also been discussed that inconsistencies in the order of the markers may be attributable to duplications, translocations, inversions or movement of transposable elements (Fu & Dooner 2002). The markers observed on the interval region of the linkage map but located outside the region of the physical map (from marker ADL0241 [~17Mbp; 80cM] to marker MCW0251 [~19.25 Mbp; 87cM]) were also explored, although none of the genes on these regions were differentially expressed.

Marker	cM	Mbp
UMA4.025	119	37.05
ADL0194	118	38.16
ADL0246	112	35.24
MCW0302	109	32.24
MCW0005	101	31.13
UMA4.046	99	28.62
MCW0251	87	19.25
MCW0114	82	18.24
ADL0241	80	17.04
LEI0100	80	17.93
LEI0095	78	25.21
ADL0145	76	17.9

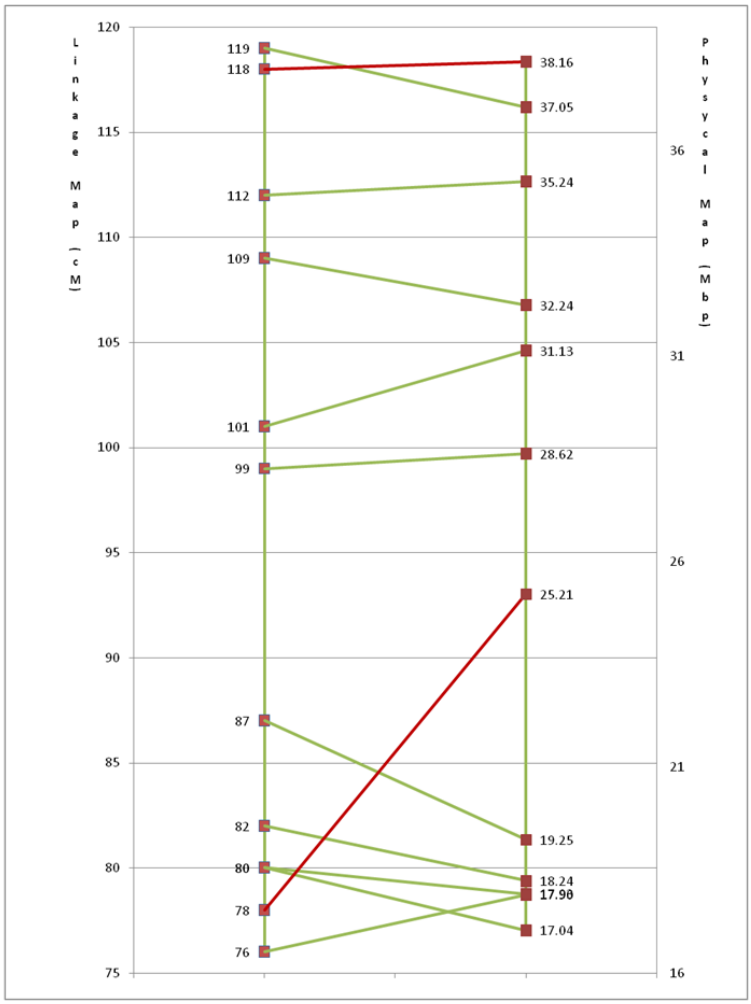


Figure 5.6 Consensus Linkage and physical map integration (QTL region)

Localization of common markers between the linkage consensus map and the physical map of the body weight QTL region. Red lines represent the closest common markers to the QTL physical flanking markers used in the study.

The location of the QTL on the consensus map allowed the utilization of the ChickenQTLdb in order to search for QTLs reported on the same region. As a result, nine other QTLs were observed (Figure 5.7). Interestingly, five of the reported QTLs are related to weight, three directly recognized as body weight QTLs (12-87 cM), one to abdominal fat weight (75 -112), and another to thigh weight (79-82 cM). The observed QTLs have an overlapping region

from the 79cM to the 82cM, and if we concentrate specially on those QTLs related to body weight the overlapping region covers 79cM to 87cM. These overlapping regions are the locations where most marker crossovers are observed (Figure 5.6).

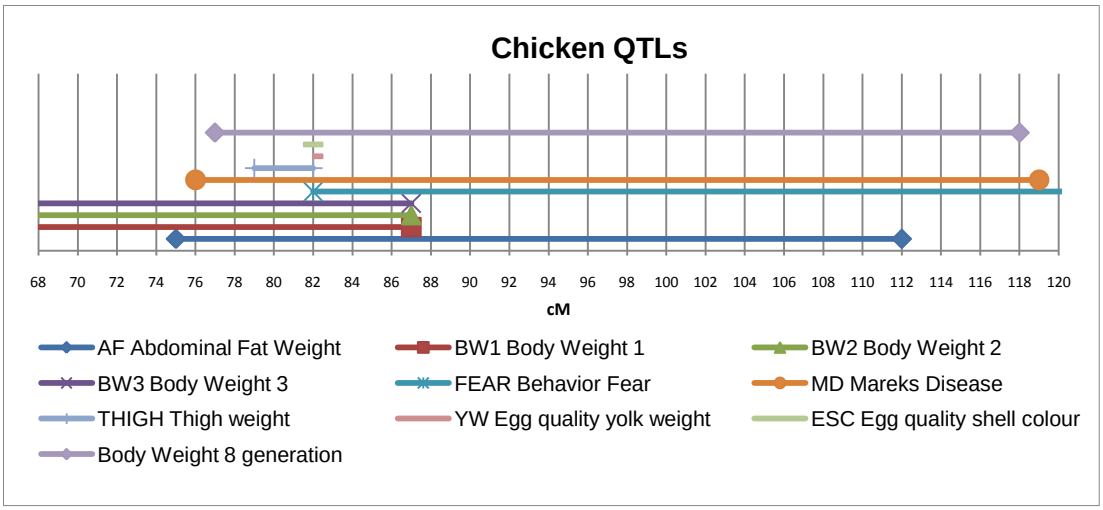


Figure 5.7 Cross QTL studies on targeted region.

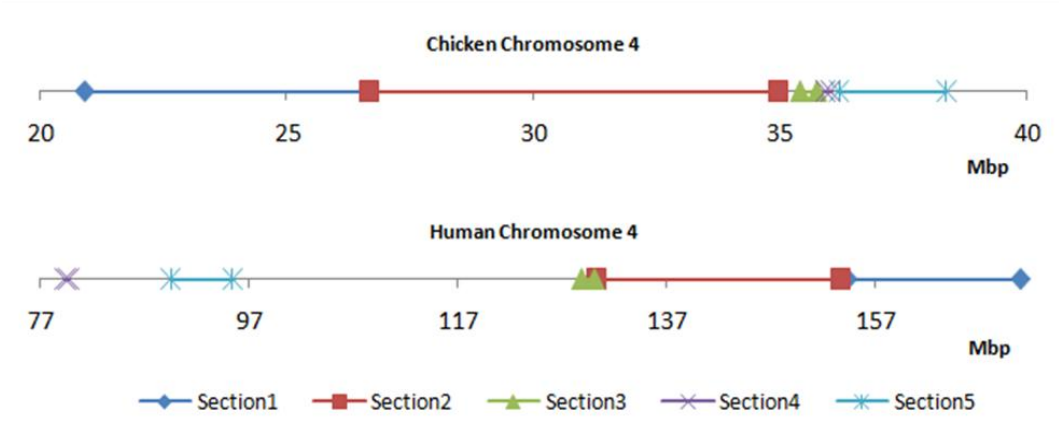
QTLs overlapping the region of the 8th generation body weight QTL (Sewalem *et al.* 2002)

5.3.1.3 Synteny Regions

The investigation of the syntenic regions of the QTL on other species allows the comparison and analysis of previously reported QTLs on the other species. At present, the comparisons between species are not straightforward. In order to obtain rat, mouse and cow synteny regions, one must obtain first the human synteny regions.

The body weight QTL region on GGA4 mapped in one human chromosome (chromosome IV), in five different sections (Figure 5.8). Each section was investigated independently in the other organisms (Table 5.1). The synteny section analysis of human and their comparison to other species showed that

some of the cow, rat and mouse syntenic sections map on various 'regions' in the chicken. This is caused by the two-step process of comparing syntenic regions across species (i.e. 1st step, from chicken to human; and 2nd from human to other species). The two-step process created overlap on some of the regions. The only implication is that some regions are counted twice and annotated to different 'sections', although their genomic coordinates and QTL annotations remain the same.



Sections	Chicken Start	Chicken End	Human Start	Human End
1	20,878,529	26,636,543	154,344,702	170,911,580
2	26,665,829	34,963,435	130,321,098	153,629,212
3	35,394,057	35,741,316	128,783,707	130,015,848
4	35,903,018	36,024,893	79,353,631	79,681,477
5	36,178,682	38,363,129	89,523,293	95,442,436

Figure 5.8 Human Synteny Regions.

QTL Human synteny regions recognized by Ensembl; resulting in five different sections on chromosome 4.

Table 5.1 Synteny Regions of Cow, and Rat.
Results of the synteny regions localized through Ensembl. *Overlap synteny regions, counted twice. (S) Chicken syntenic section

Organism	S	Chr	Start	End	QTLs
Chicken	1	4	20,878,529	26,636,543	
Cow*	1	17	1,988,015	37,695,473	Bone quality, Lean to Fat Ratio, Marbling Score, Rump Anlge, Calving Ease, Udder Height

Cow	1	17	37,699,107	46,397,033	Calving Ease, Energy Yield, Fat Percentage, Fat Yield, Milk Yield, Protein Yield, Udder Height
Cow	1	6	14,704	3,225,184	Somatic Cell Count
Cow	1	17	142,444	1,968,385	Rump Angle
Cow	1	8	185,710	7,351,489	Marbling Score, Rump Angle
Rat*	1	2	166,913,460	178,969,908	Body Weight, Blood Pressure, Prepulse Inhibition, Cardiac Mass, Non-insulin Diabetes mellitus, Renal Function, Heart Rate
Rat	1	16	23,969,429	54,259,566	Non-insulin Diabetes mellitus, Blood Pressure, Hepatocarcinoma Susceptibility, Alcohol Consumption, Aerobic Running Capacity, Consumption level Saccharin preference
Chicken	2	4	26,665,829	34,963,435	
Cow*	2	17	1,988,015	37,695,473	Bone quality, Lean to Fat Ratio, Marbling Score, Rump Anlge, Calving Ease, Udder Height
Rat*	2	2	122,858,770	141,067,643	Renal Function, Blood Pressure, Collagen Induced Arthritis, Cardiac Mass, Prostate Cancer Resistant, Smooth Muscle Cell Number
Rat	2	19	26,345,404	34,414,577	Urinary Albumin Excretion, Prostate Cancer Resistant, Non-insulin Diabetes mellitus, Kidney Mass, Tongue Tumor Susceptibility, Bone Mineral Density
Rat*	2	2	166,913,460	178,969,908	Blood Pressure, Prepulse Inhibition, Cardiac Mass, Renal Function, Body Weight, Heart Rate, Non-insulin Diabetes mellitus
Chicken	3	4	35,394,057	35,741,316	
Cow*	3	17	1,988,015	37,695,473	Bone quality, Lean to Fat Ratio, Marbling Score, Rump Anlge, Calving Ease, Udder Height
Rat*	3	2	122,858,770	141,067,643	Renal Function, Blood Pressure, Collagen Induced Arthritis, Cardiac Mass, Prostate Cancer Resistant, Smooth Muscle Cell Number
Chicken	4	4	35,903,018	36,024,893	
Cow	4	6	70,269,944	106,610,460	Fat Percentage, Milk Yield, Fat Yield, Calving Ease, Udder Attachment, Hip Height
Rat	4	14	6,338,726	37,143,140	Body Weight, Serum Renin Concentration
Rat	4	14	6,338,726	37,143,140	Non-insulin Diabetes mellitus, Renal Function, Body Weight, Blood Pressure, Serum Renin Concentration, Mammary Tumor Susceptibility, Pristane Induced Arthritis, Tongue Tumor Susceptibility, Acute Phase Response
Chicken	5	4	36,178,682	38,363,129	
Cow	5	6	3,338,006	37,748,001	Body Depth, Tenderness Score, Carcase Weight, Teat Placement, Udder Depth, Average Daily Gain
Rat	5	4	87,715,635	94,660,489	Heart Rate, Blood Pressure, Anxiety Response, Spike Wave Discharge Measurment, Pristane Induced Arthritis, Insulin Diabetes Dependent Mellitus, Uvea Inflammation Score, Serum Triglyceride Level, Lipid Level, Alcohol Consumption, Joint/Bone Inflammation, Glucose Level, Tongue Tumor Susceptibility

The RGD database was queried to obtain the human, rat and mouse body weight QTLs. In total, 495 BW QTLs were annotated to human, 105 BW QTLs to rat and 108 BW QTLs were annotated to mouse. Mouse QTLs did not have genomic annotations in the RGD database. Therefore mouse QTLs and synteny regions could not be compared to the chicken BW QTL under study. Twenty-four body weight QTLs were found on human chromosome 4, from which 8 of them fall in the synteny regions of the chicken body weight QTL under study. In rat, a total of 11 out of 15 BW QTLs were found annotated across four chromosomes mapping to the chicken QTL synteny regions (Table 5.2).

The results of the comparative and integrative analyses performed on the positional candidate genes (AADAT, JADE1 (PHF17), and INTU) found these genes also mapping on the human, rat, mouse, and cow synteny regions. The localization of the syntenic regions of the candidate genes in the various species, allowed to investigate the QTLs that have been reported for these genetic regions.

In chicken, AADAT maps to chromosome 4:26,308,672 - 26,309,204 bp. However this gene coordinates fall clearly inside chicken section 1 (section 1: 20,878,529 - 26,636,543 bp); in human, synteny section 1 is annotated to 154,344,702 - 170,911,580 bp (Figure 5.8). The human AADAT maps to chromosome 4:171,217,948 – 171,247,947 bp, locating slightly outside the annotated ENSEMBL chicken-human syntenic region. The human genomic region of this gene has been annotated to 15 QTLS; 5 of these were annotated as BW QTLs (BW79_H, BW224_H, BW441_H, BW389_H, BW379_H) (Twigger *et al.* 2007). In addition, 2 of the BW QTLs reported (BW79_H chr4: 145,556,329 - 171,556,329 bp and BW224_H chr4: 155,953,024 - 181,953,024 bp) overlap with the chicken synteny section 1. In the other species, AADAT was

found on the synteny regions of rat chr16: 32845004 – 32885600 bp; mouse chr8: 62984921 – 63024474 bp; and cow chr8: 1,677,233 – 1,701,379 bp. Although 16 QTLs are found in the same region of rat AADAT, none of those are related to body weight. AADAT rat QTLs were annotated to kidney mass, blood pressure and non-insulin dependent diabetes mellitus QTLs. According to ENSEMBL annotation the cow QTLs in the AADAT region correspond to marbling and rump angle. The AnimalQTLdb database (Hu & Reecy 2007) registered 3 QTLs in the same region (Fat thickness #2546, Fat thickness #2548 and Marbling Score #2547).

The genomic coordinates of INTU and JADE1 mapped into the syntenic section 3 (chicken chr4: 35,394,057 - 35,741,316). INTU is annotated to the chicken chr4: ~ 35,394,662- 35,395,021; and JADE1 to the chicken chr4: ~35,729,046 – 35,728,212. The identification of QTLs across species of these 2 genes was performed through genomic analyses of the syntenic region 3. In human 10 QTLs were identified in the syntenic region 3. Two QTLs are annotated as body weight QTLs (BW127_H chr4: 107,460,309 – 133,460,309 bp and BW197_H chr4: 128,971,005 – 154,971,005). The other QTLs identified on this section refer to heart rate, chronic airflow obstruction, lipid level, and joint/bone inflammation QTLs. INTU and JADE1 rat synteny regions contain a total of 46 QTLs reported, from which two are BW QTLs (BW49 and BW50) although they are annotated to the same coordinates (rat chr2: 24,474,676-163,154,227 bp). Blood pressure, bone mineral density, bone structure and strength, cardiac mass, collagen induced arthritis, glucose level, kidney mass, saccharin preference, serum cholesterol level, serum leptin concentration, smooth muscle cell number and stress response are among the annotations of the other QTLs reported to the synteny section 3 in rat. The syntenic region of chicken-cow for the section 3 maps to chromosome 17. QTLs annotated for

this region correspond to bone quality, lean to fat ratio, marbling score, rump angle, calving ease, and udder height. The QTLs on the synteny regions of other species can give us an idea if functions of sequences across species behave similarly.

Table 5.2 Human and Rat Body Weight QTLs.

Body weight QTLs found on human and rat. The first section shows the total number of QTLs found in those species, the average per chromosome and the number of QTLs falling in the synteny regions of the chicken BW QTL under study from the total annotated to those chromosomes (i.e. 4 QTLs out of 5 on rat chromosome II are in the same region of the chicken BW QTL). The second section describes the QTLs annotated to the same region of the QTL of interest. Abbreviations: Body Weight (BW), Body Fat (BF).

Species	QTLs	QTL Average x Chromosome	In QTL/Total (Chromosome)	QTL
Human	495	20.6	8/24 (IV)	8
Rat	105	5.5	4/5 (II); 4/6 (IV); 2/3 (XIV); 1/1 (XIX)	11

Species	RGD ID	Trait	Sub Trait	Chr	Start	Stop	Synteny Section
Human	1559360	BW	body mass index	4	1.46E+08	1.72E+08	1
Human	1643324	BW	body mass index	4	1.56E+08	1.82E+08	1
Human	1643249	BW		4	1.31E+08	1.57E+08	2
Human	1643255	BW	body mass index	4	1.31E+08	1.57E+08	2
Human	1643256	BW	body mass index	4	1.07E+08	1.33E+08	3
Human	1643293	BW	body mass index	4	1.29E+08	1.55E+08	3
Human	2289199	BF	percent fat	4	54152317	80152317	4
Human	2289421	BW	body mass index	4	27221423	86527489	4
Rat	1302793	BW		2	1.46E+08	1.91E+08	1,2
Rat	1358900	BW		2	1.63E+08	2.27E+08	1,2
Rat	1358887	BW		2	24474676	1.63E+08	2,3
Rat	1358908	BW		2	24474676	1.63E+08	2,3
Rat	70167	BW		4	75732943	1.19E+08	5
Rat	1357342	BW		4	75732943	1.19E+08	5
Rat	1549839	BW		4	60262965	1.17E+08	5
Rat	1549843	BW		4	60262965	1.04E+08	5
Rat	1331740	BW		14	4895894	33040140	4
Rat	631212	BF	Retroperitoneal	14	12386683	32584630	4
Rat	1354633	BW	post-adrenalectomy	19	25548711	39854409	2

5.3.2 Global Analyses

5.3.2.1 Pathways and Gene Ontologies

The results of the gene ontologies can give us an idea of the functional terms that have been associated with the list of the differentially expressed genes. The results obtained through Ontologizer (Figure 5.9) and AgBase suggested various changes under the cellular processes and metabolic process. AgBase allowed a more interpretative 'slim' annotation of the gene ontologies, the genes were grouped into their biological process, cellular component, and molecular functions.

Although GenMapp results were consistent with those found with Ontologizer and AgBase; in this study, the use of GenMapp was very advantageous. GenMapp allowed the visualization of the continuous expression changes of the genes found on the microarray through the observation of gene expression according to various FDR ranges, instead of investigating only the genes under a certain threshold.

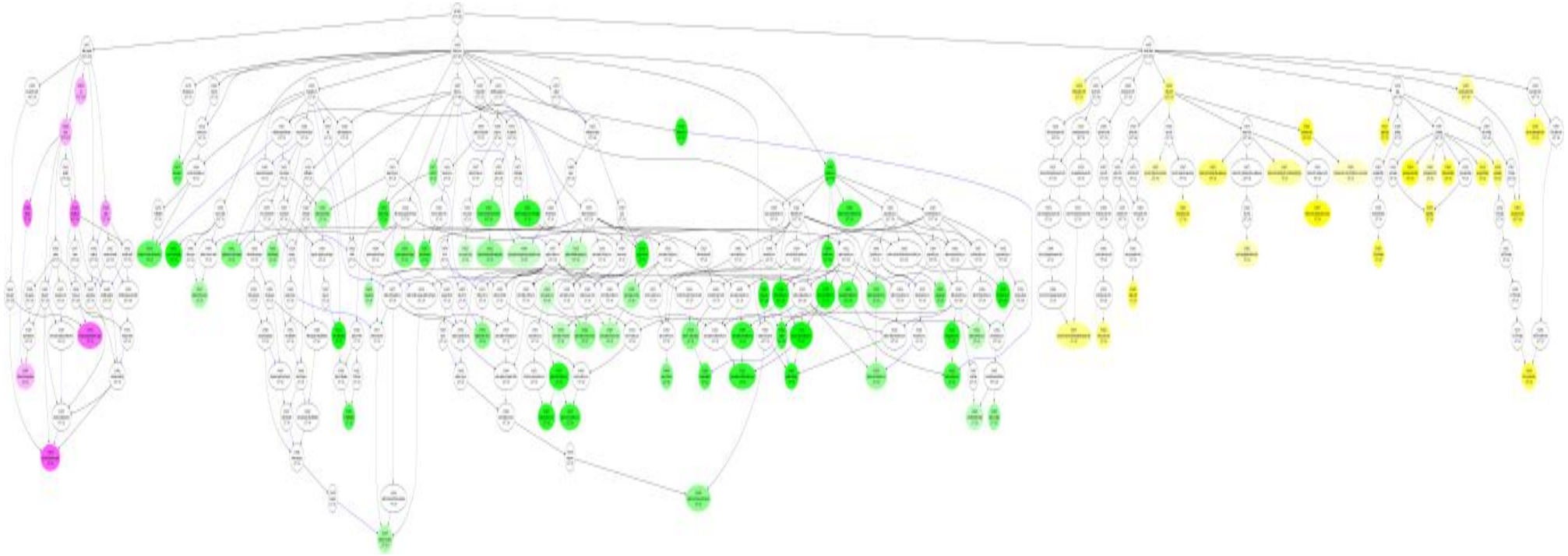


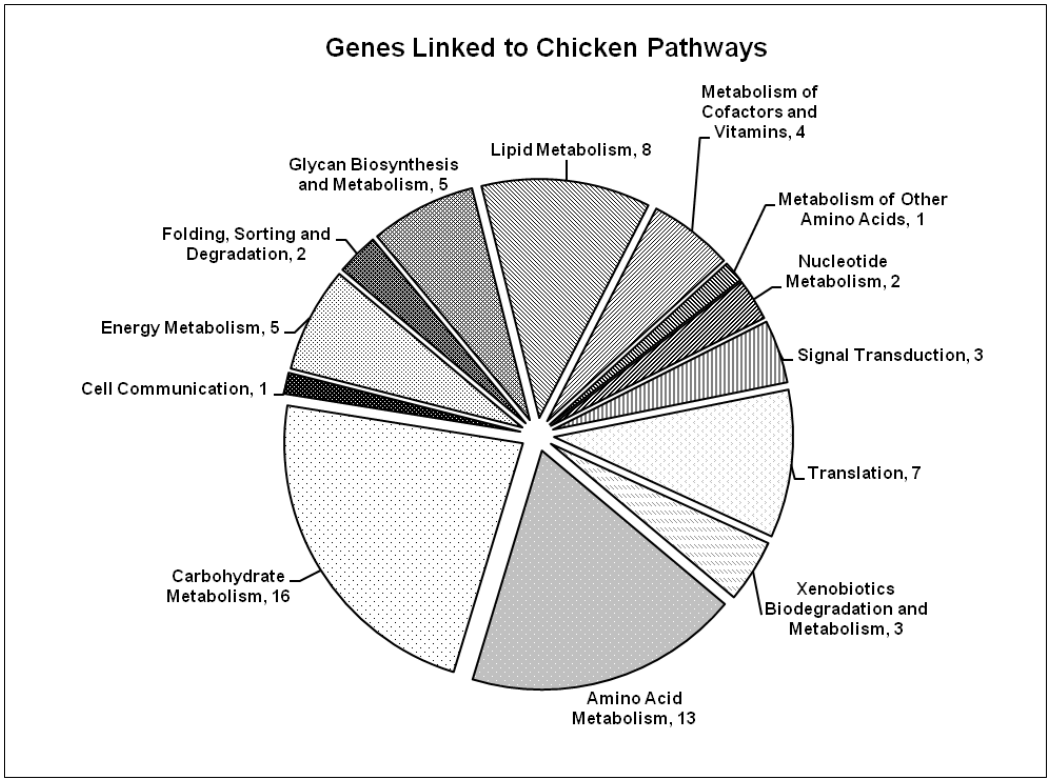
Figure 5.9 Ontologizer Results.
Gene ontology results of the genes showing differentially expression at 30% FDR. (Purple –*Cellular Component; Green – Biological Process; Yellow – Molecular Functions;* Note: original picture better viewed in Appendix 5.2 electronic version).

Pathway analyses results are difficult to interpret especially when the number of pathways and genes increases. The characterized differentially expressed genes mainly take part in the amino acid metabolism and the carbohydrate metabolism modules. More specifically, some of the significant genes participate in glycolysis and gluconeogenesis, lysine degradation, lysine biosynthesis and tryptophan pathways.

In total, chicken pathway analyses identified 40 genes with an FDR < 0.30 linked to 45 pathways (i.e. 11.2 % of the differentially expressed genes were successfully linked to pathways). As expected, some genes participate in more than one pathway. For example *alcohol dehydrogenase 5* (ADH5) acts in nine different pathways. Two other differentially expressed genes had a role in four different pathways each, while 23 genes were involved in single modules.

The global analysis (chicken data) showed the highest results in the carbohydrate metabolism, amino acid metabolism, lipid metabolism, translation and energy metabolism pathways (Appendix 5.3). The gene ontology analyses showed enrichment of oxidoreductase activity, regulation of translation, transferase activity, ubiquinone metabolic process, regulation of lipid metabolic process, among others.

Results from analyzing the data using human homologies were consistent with the ones found in chicken: the same pathways were enriched for differentially expressed genes, but with an increase of the number of significant genes in the pathways (Figure 5.10). Additionally, there was a notable concentration of differentially expressed genes in cancer and immunology human reference pathways although this could be simply a reflection of the vast amount of studies made on these areas.



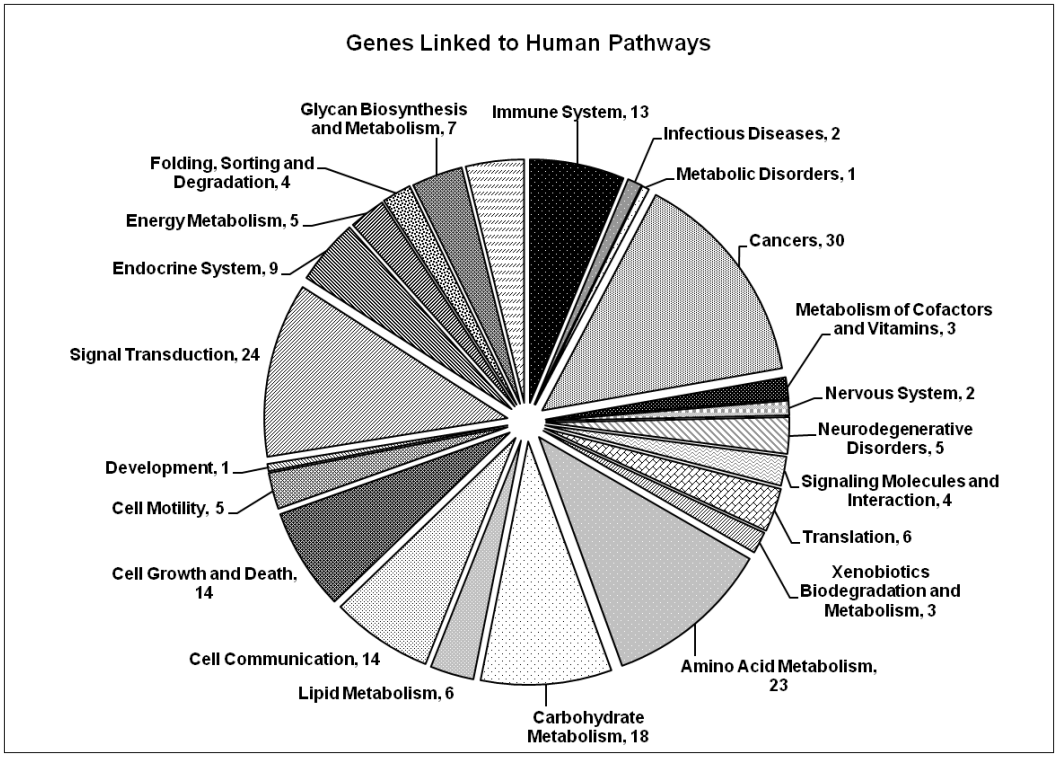


Figure 5.10 Genes linked to Pathways

Distribution of the significant genes on main pathway groups, followed by the number of genes taking part on each module. (Above) Significant genes mapped to chicken pathways; (Below) Significant genes linked to human pathways.

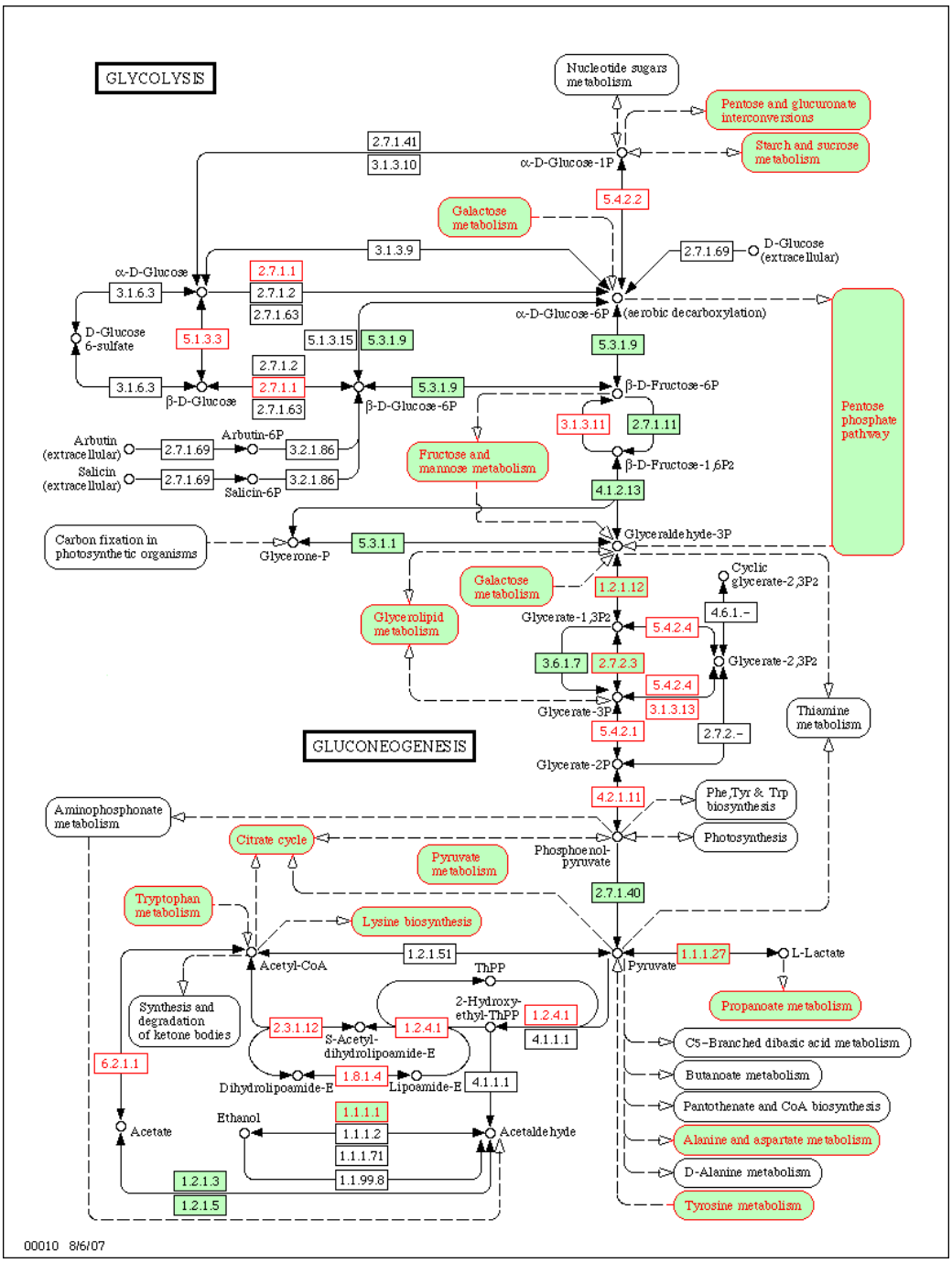
Modular Changes

An interesting finding in this study was that the expression changes through the modules and pathways tended to behave according to their genotype, showing that the modular gene expressions tend to be regulated in similar ways or by gene expression synchronization, finding changes by functional blocks. Modular changes were found and discussed by Ihmels *et al.*, (2002). For instance, for the energy metabolism module on this study all the significant genes that were identified to have a role in this module were up-regulated for the same genotype (*QQ*). The differentially expressed genes within the energy metabolism module were found in the oxidative phosphorylation, methane metabolism and carbon fixation pathways. Additionally, genes involved in energy metabolism have been related to obesity, where there is an imbalance between the intake and expenditure of energy (Labib 2003).

The carbohydrate metabolism pathway contained a high accumulation of expression changes due to the participation of two genes (LDHA and LOC418170) involved in multiple pathways. However, within the carbohydrate metabolism pathway, the glycolysis and gluconeogenesis module integrates numerous differentially expressed genes and differentially expressed pathways (Figure 5.11).

The amino acid metabolism module is composed of pathways such as lysine biosynthesis, lysine degradation, cysteine metabolism, urea cycle, tryptophan metabolism, alanine and aspartate metabolism, and tyrosine metabolism. This module showed a similar pattern as the energy metabolism; all the genes participating in these pathways were up-regulated in the *QQ* genotype, with the exception of one module where both genes (LDHA, and

SULT4A1) under cysteine metabolism were up-regulated for the *qq* genotype. Lysine degradation is one of the individual pathways that exhibited enrichment for differentially expressed genes. The significant genes found for chicken pathways involved in lysine degradation are *aminoadipate-semialdehyde synthase* (AASS), *oxoglutarate dehydrogenase-like* (OGDHL), *procollagen-lysine* (PLOD2), and LOC426314 (similar to *Histone-lysine N-methyltransferase*).



Box Color	Red Fonts	Black Fonts
Green	C, M, S	C
White	C, M	

C – Mapped in Chicken
M – Present on Microarray
S – Significant

Figure 5.11 Glycolysis and Gluconeogenesis KEGG reference pathway

(http://www.genome.ad.jp/kegg/pathway.html). Red Fonts (lined) – represents characteristics present on the microarray; red fonts with green colour box, indicates differentially expressed genes and modules; the red fonts with white background were not significant characteristics. A black font means no measurements were obtained by

microarray. Green colour boxes with black fonts were found on chicken but were not found in the microarray. White background boxes with black fonts were identified in other organisms but not in chicken.

Furthermore, pathway analyses can also guide us to the identification of genes that might be functionally related to the trait under study but are not present on the microarray. Although there are no expression levels associated to the genes, four genes were considered as possible candidate genes because of the functional relatedness of the enriched pathways to the trait (oxidative phosphorylation, lysine degradation, and tryptophan metabolism) and their proximity to the QTL under study. One of these genes, LOC770879 (similar to NADH dehydrogenase) is located in the QTL region, and the other three genes (PPA2, HADH, TDO2) are approximately 3 Mbps away from the QTL. A particular highly variable behaviour was observed for the expression levels of 18 genes encoding for the orthology enzymes (EC: 1.6.5.3 ; 1.6.99.3; minimum *p*-val = 4.65E-03 and the maximum *p*-val = 0.95) of LOC770879 (similar to NADH dehydrogenase)(Appendix 5.4).

A different example was found on the in-depth analysis of the enriched pathways for the QTL region. Initially the enzyme EC 2.1.1.43 appeared to be represented on the microarray, part of the top 50 differentially expressed genes, located inside the region of the QTL, and also containing a domain SET which was associated to contribute to epigenetic mechanisms of gene regulation and to modulation of growth control (Cui *et al.* 1998). However, by analyzing the results including the clone identifiers on the pathways, the gene located in the QTL region encoding for EC 2.1.1.43 (SETD7) was not differentially expressed between the two genotypes. This enzyme showed significant results due to a highly differentially expressed gene

(ENSGALG00000013920; *p*-value = 0.00091) located on chromosome 1 which was encoding for the same enzyme.

Pathway associations can lead us to a better understanding of the molecular changes regulating complex traits, and follow the downstream effects onto other genes even though the genes lack of expression values because they were not localized on the microarray. Additionally, pathway analyses lead to the possibility of tracking the expression changes according to the phenotypes, finding a modular behaviour on the pathway expression changes.

5.4 Conclusion

The annotation procedures were further exploited through sequence analyses and associations between the linkage and physical maps.

The GGA4 body weight QTL under study co-localized with other 8 previously reported chicken QTLs, from which 5 of them are annotated to the trait type of *growth*. The growth QTLs under the region of study (~ 78 – 118 cM) overlap from the 79 to the 87 cM. The overlapping region for the 8 QTLs was found from the 79 to the 82 cM. In addition, the AADAT candidate gene maps approximately in this region (~78 cM). This region was difficult to analyse due to marker crossovers between the linkage and physical maps. The only gene differentially expressed from ~ 17 – 28 Mbp (~76 to 99 cM) was AADAT. Furthermore, a '*golden path gap*' (no contigs detected in this part of the chromosome) was found in the physical map from the ~19.32 - 20.78 Mbps. In the linkage map, a gap with no markers was also observed between the ~87 to the 99 cM. The other two positional candidate genes, INTU and PHF17 localize ~ at the 112 cM, the QTLs found under this region refer to abdominal fat weight, fear and MDs QTLs. In addition, creatine kinase

concentration QTL (CREAT; under physiological disorders trait types) was also registered in the AnimalQTLdb at the position 82 – 138 cM (QTL centre location at 110cM). This QTL was not discussed previously. Blood creatine kinase concentration levels have been used as indicators of skeletal muscle abnormalities. Only four CREAT QTLs have been mapped across the chicken genome.

The detection of the candidate genes on other species allowed cross-species QTL comparisons. The chicken-human syntenic region of AADAT was associated with a total of 15 QTLs, five corresponding to BW. The rest of the QTLs localized in this region are annotated as heart rate, apolipoprotein, reversible airflow obstruction, myocardial infarction, and chronic airflow obstruction QTLs. Rat-chicken syntenic regions of AADAT were associated with 16 QTLs, although none of them related to body weight, these associations were related to kidney mass, blood pressure, and non-insulin dependent diabetes mellitus QTLs. Cattle QTLs in this region are linked to meat quality traits (marbling, fat thickness) and to rump angle.

The QTL results of human INTU-PHF17 region were not very variable from AADAT-human region. This could be simply due the two synteny sections map relatively close on the same chromosome (chr4). However, 10 QTLs were found in the INTU-PHF17 region, two referring to BW. In addition, lipid level, and joint/bone inflammation QTLs were also found in this syntenic region. In total, four body weight QTLs out of the eight QTLs mapped to human syntenic regions of chicken co-localize with the three positional candidate genes discussed above. The number of linked rat QTLs in the INTU-PHF17 region increased to 46, and only two related to BW. The others linked similarly to the results found in human; to blood pressure, bone mineral density, bone structure, collagen, glucose level, saccharin

preference, serum cholesterol, serum leptin, smooth muscle and stress response QTLs. Cattle results were also consistent, the QTLs found in this region were linked to bone quality, lean to fat ratio, marbling score, rump angle and calving ease.

The majority of the QTLs across species, found in the regions of the candidate genes, are mainly related to *cardio* and *growth* traits. These results supports the study made on an F_2 population cross between broilers and layers where they identified 11 genome-wide QTLs that might be associated to cardiopulmonary or muscular disorders (Navarro *et al.* 2005). The CREAT QTL previously mentioned was one of the genome-wide QTLs results they found in their study.

The gene ontologies and pathways analyses suggested changes in the cellular and metabolic process. The amino acid, carbohydrate and lipid metabolisms modules showed the highest concentration of differentially expressed genes. The AADAT was related to lysine degradation, lysine biosynthesis and trytophan metabolism pathways, all of these pathways form part of the amino acid module. Moreover, the activity of the rat and mouse *kynurenine aminotransferase II* (homodimer protein highly similar to the protein which human AADAT encodes for) consists in the transamination of alpha-aminoadipic acid. This is the final step in the major pathway (the saccaropine pathway) for the catabolism of L-lysine (AADAT NCBI reference). These homology genes also contain the *aminotransferase class I and II* conserved domain. The domain that showed disruption after the *in-silico* sequence modifications of the alternative splicing variants. INTU and PHF17 were not annotated to any pathway.

The differentially expressed genes seem to influence energy metabolism related pathways and mitochondrial cellular component genes. Based on these results it was hypothesized that the QTL on GGA4 increases body weight by increased activity of the lysine biosynthesis and degradation pathways through differentially expressed genes under the QTL.

The integration of various methodologies, together with the analysis of gene expression variations through molecular pathways, made it possible to formulate a new hypothesis for the characterization of a functional known QTL.

Appendix 5.1 In-depth Analysis

Transcripts differentially expressed in the QTL region

1) BU465968

Database Information:

BResults	RIdentifier	Status	Genbank	UniGene
0.306990684	ChEST270G22	OK	BU465968	NULL

Ensembl	LocusID	Start	End
---	NULL	26308672	26309204

TIGR_TC TC description
TC202629 ---
Component_GO ---
Function_GO ---
Process_GO ---
IPI Description ---
Protein ID --- Gene Name ---
Unigene Description ---
Gene --- Locuslink ---
Similar to Human Molecule
Similar to Mouse Molecule
Similar to Rat Molecule
Blast Match Unigene Representative Sequence Description

The sequence was extracted from NCBI-Nucleotide (30 May 2007) and Blast in NCBI and ENSEMBL

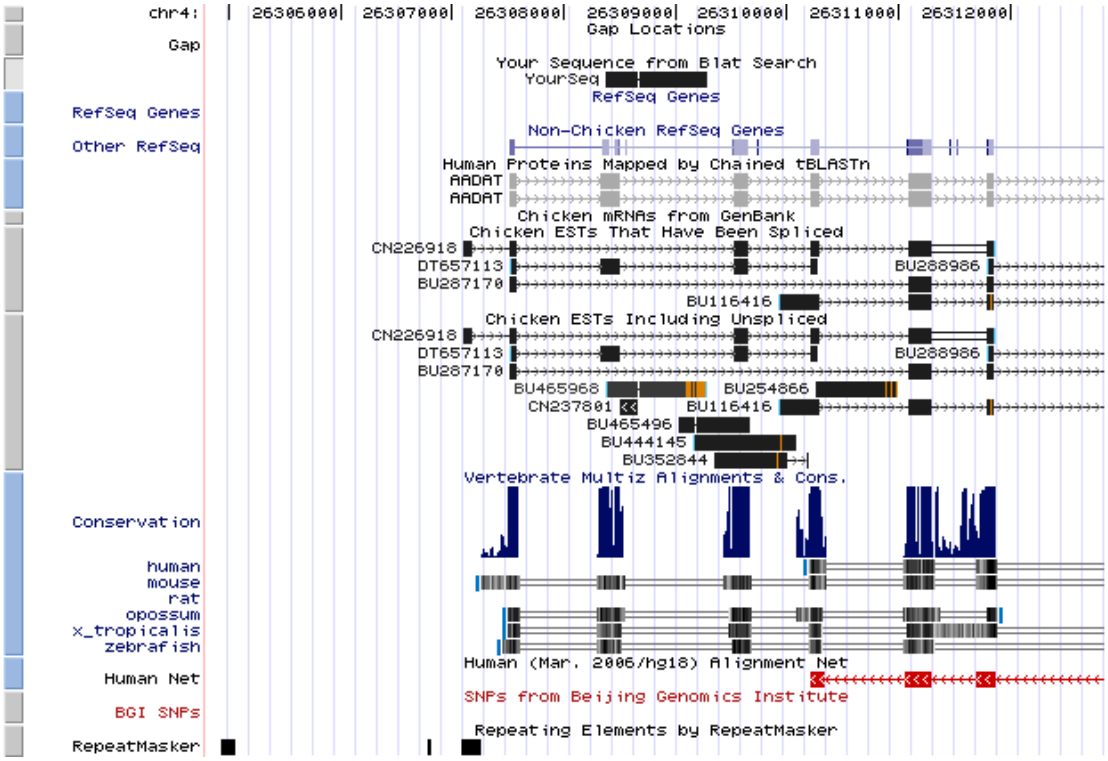
Nucleotide sequence of ChEST270g22:

>gi|25955442|gb|BU465968.1|BU465968 603368084F1 CSEQRBN19 Gallus gallus cDNA clone ChEST270g22 5', mRNA sequence
TACGGGGCACCAAACCCTGCCGTTTTTCCATTTAAGAAGGCTACTATTGCCACTGGACATGGAAATGCTG
TTGAGATTGGGGAAGACTTAATGAAGAGGGCTCTTCAATACTCTGCCTCAGCAGGGTATCTGTCAAGGCT
GCTTGCACTAAAAGATGGTGGAACAAAAGAATCCTGAAGTTCTCTGTCATTGCATTAATTGATCTTTACA
AAAACAGCTTTAACTTCTTTAGATTTTCCTCAGTGTTTATATTTCAGTTTAAGCAGAAGGTTCTGAGTAG
ATGAGCATTTTTTTTTGGACAAGCCAGAAATTTCCCAGTATCTGAGAATCCCAACTCCCAAGTTAGTAAA
AGGACTCATCCTGTATTTGTATTAAGATATTCAGTGAATGATGTTGTATTGAAGTGTGGTTTCCAGCTCC
TGAAGCATTCATTTTGCCTTTATTTGGTACCCTAAAGAAGTCGACACAGACCACTTGTGGTGCTCGAACT
GAGGTTTTTATAGTATTGGCATTTTTACTACAAAAATTGCATATGTTATGTTAGTTTAGCACTCTCAGTA
AGTCCCCAGTAGGCAGCACCAAGTCAGGCTTATGACAGGCTTGCAAAAATACGCGTGTACCTCCATAGTC
ACACAGGGCATTAATCAGTGCAGTTGCAGCTTTTCCTAATGCTTAAGCGAACACAACTTTAGTAAAAGGC
ATTTGTAGACGTAACTACAAGACTTTTATCTTAAAACAGCTCTTTTGAAAAACTCATTTTCTGAATTTCT
TGCTAGCTGCAGGCCCACTTCCAGAAGCGCTGGCAGCCTTTGTCAGGGCAGCCACATCTGAGTATTGGGA
AAGGCTTGGTGCCTAGGAAGCAGATCCTGTGGAACCTAACACCACGTTCCATCCCTCAGTGGAACCCATT
CTGGCCGATTTCCCAGTCACCG

UCSC:

BLAT Search Results

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
826	4	906	932	95.8%	4	+	26308376	26309268	893



NCBI:

>ref|NW_001471683.1|Gga4_WGA109_2 Gallus gallus chromosome 4 genomic contig, reference assembly
(based on Gallus_gallus-2.1)
Length=17760008

Features in this part of subject sequence:
hypothetical protein

ENSEMBL

Name	Start	End	Ori	Name	Start	End	Score	E-val	%ID	Length
ENSGALT00000015723	117	239	+	Chr:4	26308376	26309523	119	5.7e-61	99.19	123

Transcript:LOC428728 Ensembl Transcript ID:ENSGALT00000015723
Transcript Information :
Exons: 13 Transcript length: 1,535 bps Translation length: 428 residues
This transcript is a product of gene: ENSGALG00000009655
This transcript can be found on Chromosome 4 at location 26,307,508-26,321,745.

The following GO terms have been mapped to this entry via UniProt and/or RefSeq:
GO:0005739 [from] [mitochondrion] IEA
IPR004839 Aminotransferase, class I and II

ENSF00000002471 : KYNURENINE/ALPHA AMINOADIPATE AMINOTRANSFERASE MITOCHONDRIAL PRECURSOR KAT/AADAT KYNURENINE OXOGLUTARATE TRANSAMINASE II EC_2.6.1.7 KYNURENINE AMINOTRANSFERASE II KYNURENINE OXOGLUTARATE AMINOTRANSFERASE II 2 AMINOADIPATE TRANSAMINASE EC_2.6.1.- 39 2 AMIN
This cluster contains 1 Ensembl gene member(s) in this species.

Alignment:

```
Query:   4 GGGGCACCAAACCCTGCCGTTTTTCCATTTAAGAAGGCTACTATTGCCACTGGACATGGA 63
           ||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct: 117 GGGGCACCAAACCCTGCCGTTTTTCCGTTTAAGAAGGCTACTATTGCCACTGGACATGGA 176
```

```
Query:  64 AATGCTGTTGAGATTGGGGAAGACTTAATGAAGAGGGCTCTTCAATACTCTGCCTCAGCA 123
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 177 AATGCTGTTGAGATTGGGGAAGACTTAATGAAGAGGGCTCTTCAATACTCTGCCTCAGCA 236

Query: 124 GGG 126
           |||
Sbjct: 237 GGG 239
```

2) BU463895

Database Information:

BResults	RIdentifier	Status	Genbank	UniGene
0.291952766	ChEST711B18	OK	BU463895	NULL

Ensembl	LocusID	Start	End
---	NULL	32339279	32339853

TIGR_TC TC223225 **TC description** ---
Component_GO ---
Function_GO ---
Process_GO ---
IPI Description ---
Protein ID --- **Gene Name** ---
Unigene Description ---
Gene --- **Locuslink** ---
Similar to Human Molecule ---
Similar to Mouse Molecule ---
Similar to Rat Molecule ---
Blast Match Unigene Representative Sequence Description

The sequence was extracted from NCBI-Nucleotide (30 May 2007) and Blast in NCBI and ENSEMBL

Nucleotide sequence of ChEST711B18:

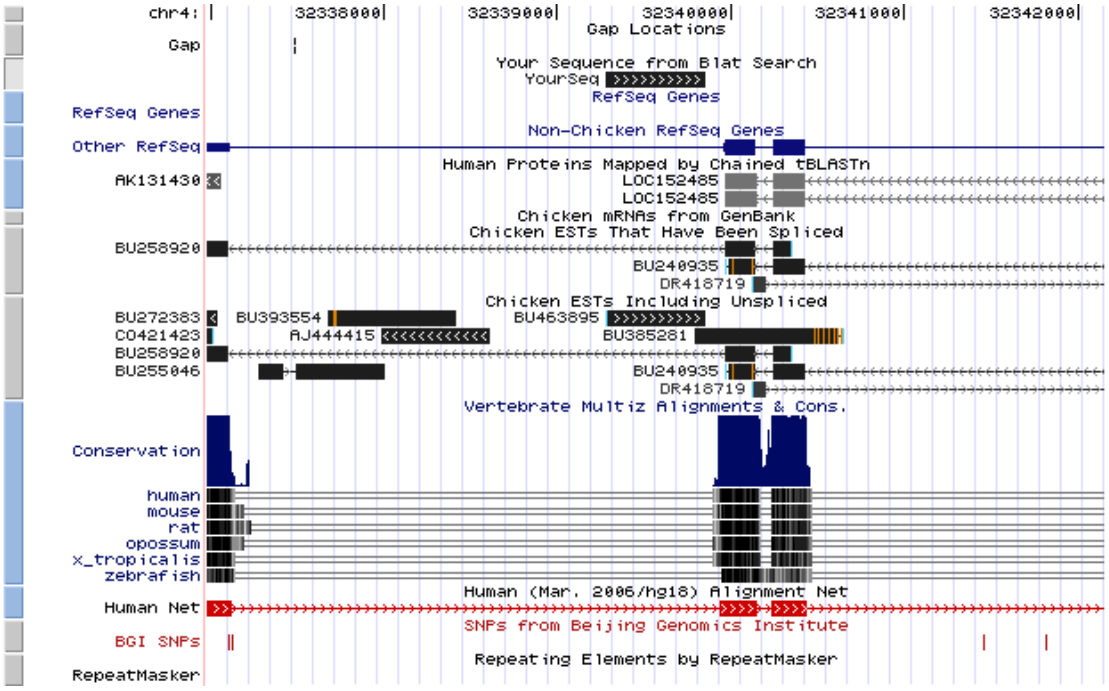
>gi|25953369|gb|BU463895.1|BU463895 603773865F1 CSEQRBN19 Gallus gallus cDNA clone ChEST711b18 5', mRNA sequence
CACGTGCTAGCCCGCCTCTCCTCAGGGGTCCAACTTTGACATTGCTCTACGAACAGATCTTGAAGGACAG
CACCTTAGCCAAAGACAGACTTTTTTTTTCTTAAAAGAGTATACTGCTTGCTACACATTAATTTATAGTG
ATTATTCCAAGCATGCAAAGAAAGGCAGTAAAGGATCAAAATAGTTAGTTGCTGAGGAGGGCTGACTTTG
CATTAGAATCTACAAGGATTACAAAGCAACTGAGAGGTCTGATTTTGAGAAGCCAACTCCATTCTTAGTT
GCACAGATTGGACAGGATTTAGATTCTTTGAGCTAGGTGCAGAACAGAGAATGTGGTGCATGCTCTCCTT
TGCTCTAATGCCTTCAAAATGCCTCTTAGTTGGACCGCACCTCCGTGCTGTACTGCCTCATTCTCTTAGG
ACCCATCACCTTCTTTCTCCTCTTTGACTTGAAATGGGAAAGGCCACACTTTGTTAAGATCTACAACTAG
GGGAAACTACATATATTACTATTTACTATAATGGTGCAAACCCCTGTTAAGGCAACTCTGAAGGACTGGG
CTTTTATTATCTAAATAACAGAAGCCCCA

UCSC:

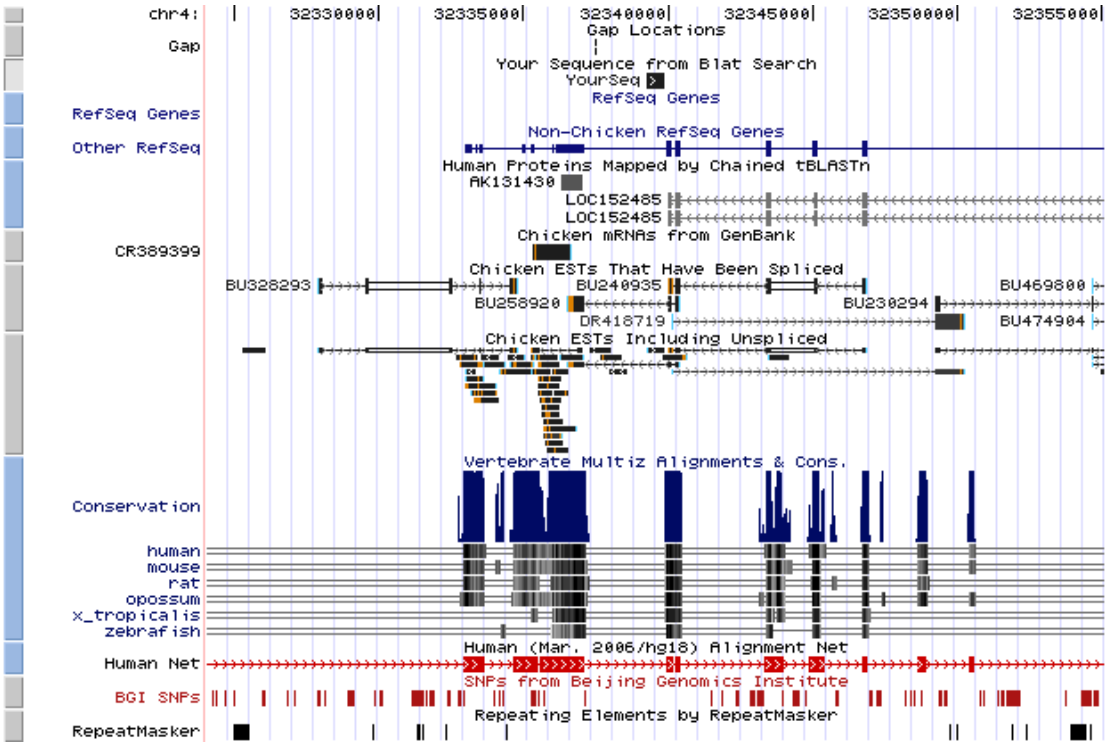
Match the region of the chromosome 4 outside any exon of the locus found in the region LOC152485

BLAT Search Results

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
575	15	589	589	100.0%	4	+	32339279	32339853	575



Zoom Out:



Near ESTs looked at:

- BU258920 (Not in the microarray)
- BU240935 (Not in the microarray)
- DR418719 (Not in the microarray)

- BU385281 (Not in the microarray)
- BU393554 (Not in the microarray)
- AJ444415 (Not in the microarray)

NCBI:

```
>ref|NW_001471683.1|Gga4_WGA109_2  Gallus gallus chromosome 4 genomic
contig, reference assembly
(based on Gallus_gallus-2.1)
Length=17760008

 Features in this part of subject sequence:
   similar to LOC152485 protein
```

ENSEMBL:

Several matches the best one hits to chromosome 10 not chromosome 4.

Name	Start	End	Ori	Name	Start	End	Ori	Score	E-val	%ID	Length
ENSGALT00000003900	695	712	+	Chr:10	336600	336617	+	18	0.67	100.00	18
ENSGALT00000012496	1115	1132	+	Chr:7	9028673	9028690	-	18	0.67	100.00	18
ENSGALT00000021016	306	325	+	Chr:2	69177546	69188949	-	17	2.6	95.24	21
ENSGALT00000032344	1273	1295	+	Chr:4	11053366	11053388	-	17	2.6	92.00	25
ENSGALT00000022069	883	899	+	W_random	44748	44764	-	17	2.6	100.00	17
ENSGALT00000012245	657	677	+	Chr:6	23758312	23758332	+	17	2.6	95.24	21
ENSGALT00000018891	606	622	+	Chr:Z	42325327	42325343	-	17	2.6	100.00	17

```
Ensembl Transcript ID: ENSGALT00000032344
Transcript information: Exons: 10 Transcript length: 1,363 bps Translation
length: 399 residues
This transcript is a product of gene: ENSGALG00000007355
This transcript can be found on Chromosome 4 at location 11,053,298-
11,060,985.
The start of this transcript is located in Contig Contig11.441.
InterPro:
       IPR007110 Immunoglobulin-like
       IPR013106 Immunoglobulin V-set
       IPR003599 Immunoglobulin subtype
       IPR013098 Immunoglobulin I-set
Protein Family: ENSF00000024051 : UNKNOWN


Query location     : gi|25953369|gb|BU463895.1|BU463895      147 to      171
(+)
Database location  : ENSGALT00000032344                     1273 to     1295
(+)
Genomic location   : 4                                   11053366 to 11053388
(-)

Alignment score    : 17
E-value            : 2.6
Alignment length   : 25
Percentage identity: 92.00
Query:  147 CCAAGCATGCAAAGAAAGGCAGTAA 171
            |||||| || ||||||||||||||
Sbjct: 1273 CCAAGC-TG-AAAGAAAGGCAGTAA 1295
```

3) BU452163

BResults	RIdentifier	Status	Genbank	UniGene
0.306281675	ChEST206D5	C	BU452163	Gga.34870

Ensembl	LocusID	Start	End
ENSGALG00000010170	422494	35394662	35395021

TIGR_TC	TC description
TC218635	"similar to UP|Q80TG0 (Q80TG0) MKIAA1284 protein (Fragment), partial (38%)"

Component_GO ---
Function_GO ---
Process_GO ---
IPI Description PREDICTED: similar to PDZ domain containing 6
Protein ID XP_420459 **Gene Name** XP_420459
Unigene Description "Transcribed locus, strongly similar to XP_420459.1 PREDICTED: similar to PDZ domain containing 6 [Gallus gallus]"
Gene --- **Locuslink** ---
Similar to Human Molecule ---
Similar to Mouse Molecule ---
Similar to Rat Molecule ---
Blast Match Unigene Representative Sequence Description ---

The sequence was extracted from NCBI-Nucleotide (31 May 2007) and Blast in NCBI and ENSEMBL

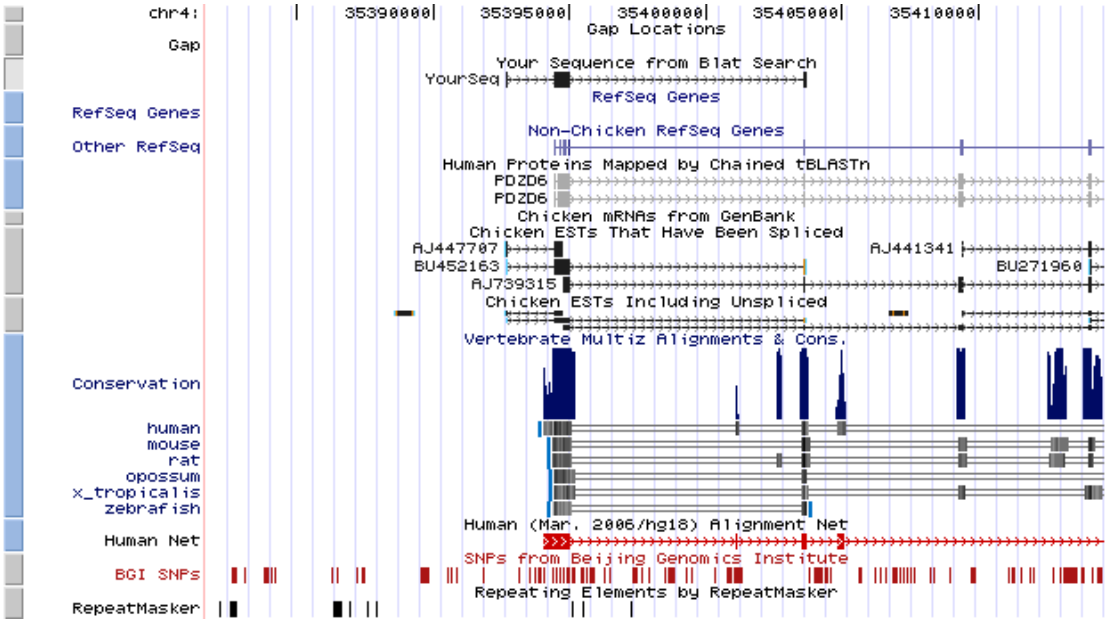
Nucleotide sequence of ChEST206D5:

>gi|25941474|gb|BU452163.1|BU452163 603217618F1 CSEQRBN14 Gallus gallus cDNA clone ChEST206d5 5', mRNA sequence
CACGGGTGTGGAGCCGCTGCAGGGTGGCAGCACGGGGTGCGCGGACACCGGCAGCGAGAGCGGCTCGGAC
AGCACCTCGCTGTCCGCCAGCTCCGATGACCTTGAGCCTGAATGGTTGGATGGTGTGCAGAAAAATGGGG
AGCTATTTTATTTAGAAATGAGTGAAAGTGAAGAAGAAACTTTACTTCAGAACGCCTGTCCAGAAATACC
ATCGGTGAATCATGTCAGATTTCGTGAAAATGAAGCGGAAGTTATTCAGGAGGGATCACGAAAAGAAAGA
AAGTATGAACTGAAGAAACTGACAAAAATCTTAAAGAAGAAGAATCTTTTACCAAAGCATTCTAGTAAGA
AAGGAAGTGGAAGCTGTAACGTGCGCTCCAGTGGTCCAACTTCCATACTGAAACACCACTCCACTCAGAA
AATGGGTGAAATACAGCAGAAGTACAAAGATATCTATGTTTATGTAAATCCCAGAAAACTGTTGGGGAAT
GCTGGAGAAGATGAGCAGCACAGGCTGCTAGAGGCCTTGGTAGGAATTCTCCATCAGTCTTCATGGAGCA
GCAGAAGAGCGGAAAAACAAGGCAAGAAGGATAAGGTCACCAGAGGAATCACTGAAGAGAAGCTTGTAGT
ACATGGCTTGGTGCCAGGCGGTTCAGCAATGAAAACAGGCCAAATATTGATTGGAGATGCTCTAGTTGCT
GTACATGATGTCGATGTGAATTCTGAAAAACATAGAAAGAGTTTGTCTTGCATTCCAGGTCCTATGCGGG
TTAAAA

UCSC:

BLAT Search Results

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
747	4	767	776	99.1%	4	+	35392667	35403652	10986



Near ESTs looked at:

- AJ447707 (Not in the microarray)
- AJ739315 (Not in the microarray)

NCBI:

>ref|XM_420459.2| PREDICTED: Gallus gallus similar to PDZ domain containing 6 (LOC422494),
mRNA
Length=3567

Score = 1371 bits (742), Expect = 0.0
Identities = 764/774 (98%), Gaps = 3/774 (0%)
Strand=Plus/Plus

ENSEMLB:

Name	Start	End	Ori	Name	Start	End	Ori	Score	E-val	%ID	Length
ENSGALT00000016536	1	715	+	Chr:4	35392724	35409280	+	677	0.	98.61	717
ENSGALT00000022249	119	136	+	Chr:4	65926031	65926048	-	18	0.79	100.00	18
ENSGALT00000001292	117	134	+	Chr:27	1279216	1279233	+	18	0.80	100.00	18

Transcript: LOC422494
Ensembl Transcript ID : ENSGALT00000016536
Transcript information: Exons: 17 Transcript length: 3,177 bps Translation length: 912 residues
This transcript is a product of gene: ENSGALG00000010170
This transcript can be found on Chromosome 4 at location 35,392,724-35,431,125.
The start of this transcript is located in Contig Contig14.125.

This Ensembl entry corresponds to the following database identifiers:
RefSeq peptide predicted: XP_420459.2 [Target %id: 97; Query %id: 95]
RefSeq DNA predicted: XM_420459.2
EntrezGene: LOC422494
InterPro IPR001478 PDZ/DHR/GLGF
Protein Family ENSF00000007954 : PDZ DOMAIN CONTAINING 6 INTURNED PLANAR CELL POLARITY EFFECTOR HOMOLOG

Query location : gi|25941474|gb|BU452163.1|BU452163 61 to 776 (+)
Database location : ENSGALT00000016536 1 to 715 (+)

```
Genomic location   : 4                                    35392724 to 35409280 (+)

Alignment score    : 677
E-value            : 0.
Alignment length   : 717
Percentage identity: 98.61
Query:  61 CGGCTCGGACAGCACCTCGCTGTCCGCCAGCTCCGATGACCTTGAGCCTGAATGGTTGGA 120
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   1 CGGCTCGGACAGCACCTCGCTGTCCGCCAGCTCCGATGACCTTGAGCCTGAATGGTTGGA 60

Query: 121 TGGTGTGCAGAAAAATGGGGAGCTATTTTATTTAGAAATGAGTGAAAGTGAAGAAGAAAC 180
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  61 TGGTGTGCAGAAAAATGGGGAGCTATTTTATTTAGAAATGAGTGAAAGTGAAGAAGAAAC 120

Query: 181 TTTACTTCAGAACGCCTGTCCAGAAATACCATCGGTGAATCATGTCAGATTTCGTGAAAA 240
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 121 TTTACTTCAGAACGCCTGTCCAGAAATACCATCGGTGAATCATGTCAGATTTCGTGAAAA 180

Query: 241 TGAAGCGGAAGTTATTCAGGAGGGATCACGAAAAGAAAGAAAGTATGAACTGAAGAAACT 300
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 181 TGAAGCGGAAGTTATTCAGGAGGGATCACGAAAAGAAAGAAAGTATGAACTGAAGAAACT 240

Query: 301 GACAAAAATCTTAAAGAAGAAGAATCTTTTACCAAAGCATTCTAGTAAGAAAGGAAGTGG 360
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 241 GACAAAAATCTTAAAGAAGAAGAATCTTTTACCAAAGCATTCTAGTAAGAAAGGAAGTGG 300

Query: 361 AAGCTGTAACGTGCGCTCCAGTGGTCCAACTTCCATACTGAAACACCACTCCACTCAGAA 420
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 301 AAGCTGTAACGTGCGCTCCAGTGGTCCAACTTCCATACTGAAACACCACTCCACTCAGAA 360

Query: 421 AATGGGTGAAATACAGCAGAAGTACAAAGATATCTATGTTTATGTAAATCCCAGAAAACT 480
           |||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 361 AATGGGTGAAATACAGCAAAAGTACAAAGATATCTATGTTTATGTAAATCCCAGAAAACT 420

Query: 481 GTTGGGGAATGCTGGAGAAGATGAGCAGCACAGGCTGCTAGAGGCCTTGGTAGGAATTCT 540
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 421 GTTGGGGAATGCTGGAGAAGATGAGCAGCACAGGCTGCTAGAGGCCTTGGTAGGAATTCT 480

Query: 541 CCATCAGTCTTCATGGAGCAGCAGAAGAGCGGAAAAACAAGGCAAGAAGGATAAGGTCAC 600
           |||||||||||| |||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct: 481 CCATCAGTCTTCGTGGAGCAGCAGAAGAGCAGAAAAACAAGGCAAGAAGGATAAGGTCAC 540

Query: 601 CAGAGGAATCACTGAAGAGAAGCTTGTAGTACATGGCTTGGTGCCAGGCGGTTCAGCAAT 660
           |||||||||||| |||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct: 541 CAGAGGAATCAGTGAAGAGAAGCTTGTAGTACATGGCTTGGTGCCTGGCGGTTCAGCAAT 600

Query: 661 GAAAACAGGCCAAATATTGATTGGAGATGCTCTAGTTGCTGTACATGATGTCGATGTGAA 720
           |||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct: 601 GAAAACAGGCCAAATATTGATTGGAGATGCTCTAGTTGCTGTAAATGATGTCGATGTGAA 660

Query: 721 TTCTGAAAAACATAGAAAGAGTTT-GTCTTGCATTCCAGGTCCTATGCGGGTTAAAA 776
           ||||||||| |||||||||||||| |||||||||||||||||||||||| ||| ||||
Sbjct: 661 TTCTGAAAA-CATAGAAAGAGTTTTGTCTTGCATTCCAGGTCCTATGCAGGT-AAAA 715
```

4) BU415609

BResults	RIdentifier	Status	Genbank	UniGene
0.237443048	ChEST606L22	C	BU415609	NULL

Ensembl	LocusID	Start	End
ENSGALG00000010214	NULL	35729046	35728212

TIGR_TC **TC description**
BU415609 ---
Component_GO ---
Function_GO ---
Process_GO ---
IPI Description
Protein ID ENSGALP00000031599 **Gene Name** ENSGALP00000031599
Unigene Description ---
Gene --- **Locuslink** ---

Similar to Human Molecule ---
Similar to Mouse Molecule ---
Similar to Rat Molecule ---
Blast Match Unigene Representative Sequence Description

The sequence was extracted from NCBI-Nucleotide (31 May 2007) and Blast in NCBI and ENSEMBL

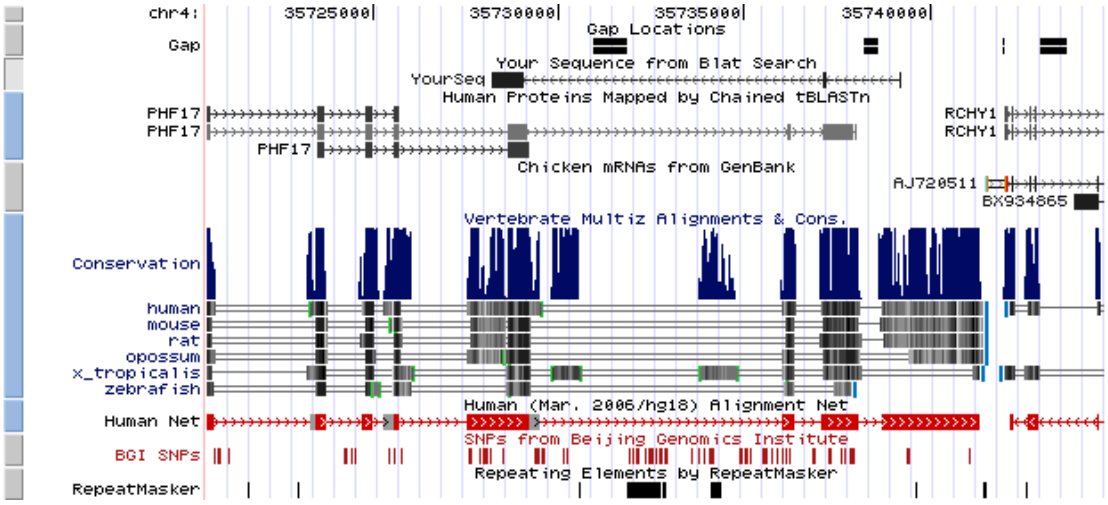
Nucleotide sequence of ChEST606L22:

>gi|25908280|gb|BU415609.1|BU415609 603667586F1 CSEQRBL06 Gallus gallus cDNA clone ChEST606l22 5', mRNA sequence
GCGAAAGAGAGGACAAAGACCTACCACCATCCAAGGGGCATTGGGACCCAAAGATGGACTATTGAGACAA
CACTGCGTTTTCCATTGTGGAGAATGAAGAAGGCACACCTAAAGCTTCCAGTACTGATAAAGGAATCCCA
CCAGCTCCTCAGGCAGCTGCAGTGCTTTAGCCACTTCCAGAGACTCGACAAAGGTATAGAATTCATCCTC
CAGCTGCTGGAGCTTCTGCTTGCGAAGACTGACTCTGTGGGCTTCCTCCTGGTTTTGATCCATGCTGTGG
AAAGGGTCAATGTGGGCAGGAAGAGAGCTGTCCTGAATCCCATTCCCATTCTCTTGACACAGGTGGTCAC
TGAAGGTCTCATCATCTGCTCTCTTGGTGGAGCTGTGCTTGGGACAGTATGACTTAAATTTCACCTCATC
GTTCTCTGCCAGTATGGTCTTCATCTCCAAGCCACGGTCAAACGCGCAAGTGACGTGAAAGGCTGTTCTG
CAGTTCTTCACTGAACACTGCAGAAAGTAGATTAAAATTCATCAAGTATAAATTTAAGAAACCACAATGC
CTAAAACTAAGCAGCCAGGACTGCTTTGAAAGAATCGACCCTGCTACCCAGCCTCTGCTTCTAAGCAGCC
CTGTACTTTTTTTGCTCTCTTTAATTAGGTTTGAGATTTAGTGCTTTCAGGCCAGGCAGCAACCAAGAGG
CCAGTGGGGACAGGAGTGGTTAACCTCTCCCTAGCCCAATCCCTCTATTCGATAATCAGCAAATCCACAA
GAGAAGTGTTCTTACCTTGACTGAAGCACTGGCTGAGAACAGTGTTACCCAAGCAAATGTCCCCCTTTTG
AATGTCAGTGTGAGACGAGAGCGACAATACATCTGTGGCGGCTCCAGGGCAGTTTGACTACTTCTATTAA
GTCTTTCCGGCACTGGGACCCTTGCTACCAAAAAGGCTCACTGATAGGTTTTTTCATCCCCTTGCCAAAG
TTATAAGGCAATTTTAATCCTGTCATCCCGGCTTCAATACTAAGGCTTCTTTCCCTGGGGTCAAACA

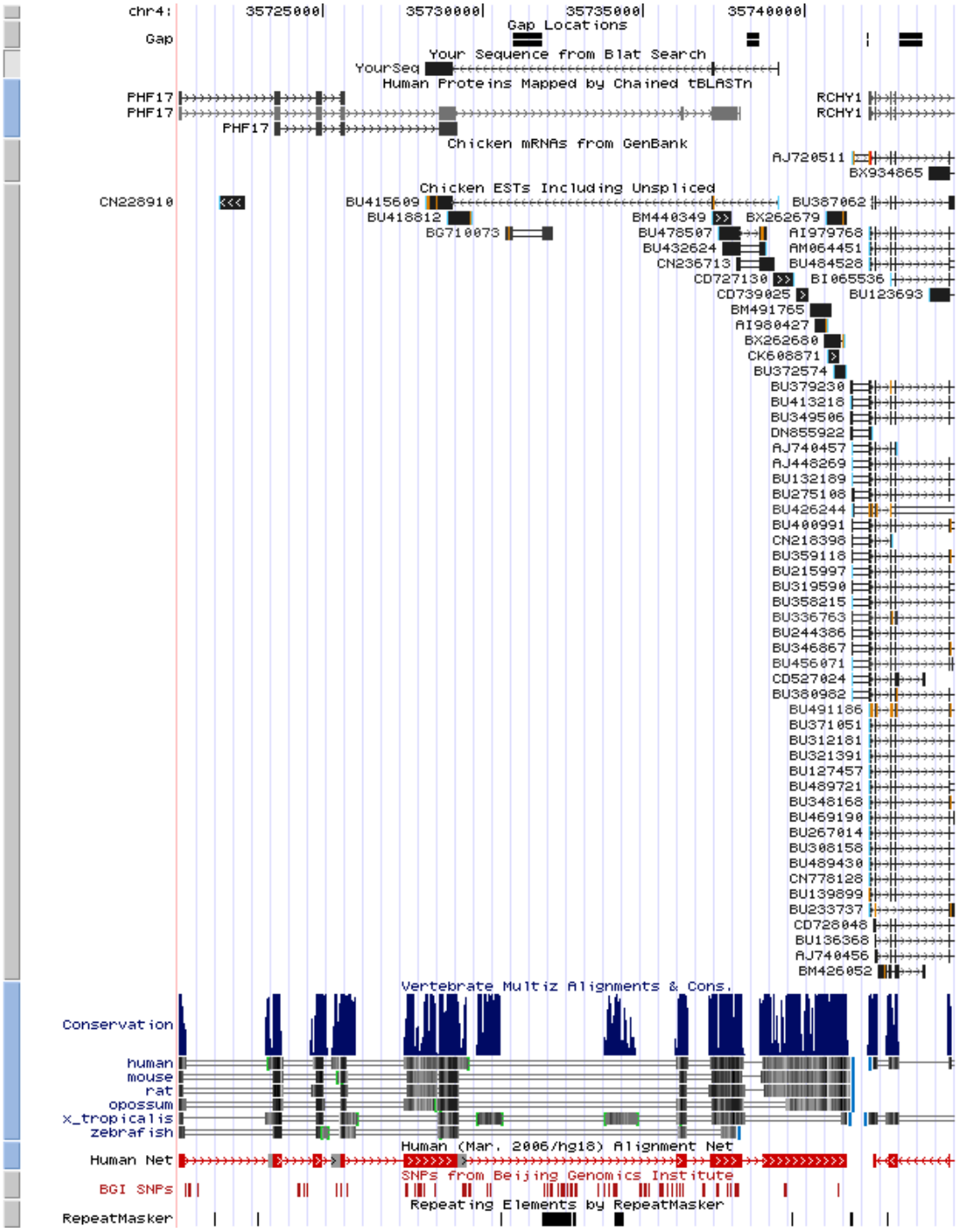
UCSC:
PHF17: Uniprot:JADE1_HUMAN:

BLAT Search Results

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
895	4	970	1047	97.1%	4	-	35728184	35739178	10995



With ESTs:



Near ESTs looked at:

- BU418812 (Not in the microarray)
- BG710073 (Not in the microarray)
- BM440349 (Not in the microarray)
- BU478507 (Not in the microarray)
- BU432624 (Not in the microarray)
- CN236713 (Not in the microarray)
- CD727130 (Not in the microarray)

NCBI:

>ref|XR_027202.1| PREDICTED: Gallus gallus similar to JADE1L protein (LOC422502),
mRNA
Length=2677
Score = 736 bits (398), Expect = 0.0
 Identities = 398/398 (100%), Gaps = 0/398 (0%)
 Strand=Plus/Minus

>ref|XM_001365262.1| PREDICTED: Monodelphis domestica similar to novel transposon
(LOC100011405), mRNA
Length=2577

 Score = 165 bits (89), Expect = 3e-37
 Identities = 306/408 (75%), Gaps = 26/408 (6%)
 Strand=Plus/Minus

ENSEMBL:

Name	Start	End	Ori	Name	Start	End	Ori	Score	E-val	%ID	Length
ENSGALT00000016607	976	1373	+	Chr:4	35725696	35729046	-	398	3.0e-266	100.0	398
ENSGALT00000016607	1626	1707	+	Chr:4	35736264	35737199	-	79	3.0e-266	98.80	83
ENSGALT00000010368	1114	1149	+	Chr:13	16195882	16195917	+	24	3.4e-07	91.67	36
ENSGALT00000019730	1646	1664	+	Chr:2	46221989	46222007	-	19	0.31	100.00	19

Transcript XR_027202.1 (RefSeq DNA)
Ensembl Transcript ID: ENSGALT00000016607
Transcript information Exons: 11 Transcript length: 2,650 bps Translation length: 847 residues
This transcript is a product of gene: ENSGALG00000010214
This transcript can be found on Chromosome 4 at location 35,703,692-35,738,758.
The start of this transcript is located in Contig Contig14.115.
Description PREDICTED: Gallus gallus similar to JADE1L protein (LOC422502), mRNA. Source: RefSeq_dna XR_027202

This Ensembl entry corresponds to the following database identifiers:
RefSeq DNA: XR_027202.1 [Target %id: 94; Query %id: 93] [align]
EntrezGene: LOC422502
UniGene: Gga.25397 [Target %id: 94; Query %id: 93]

GO The following GO terms have been mapped to this entry via UniProt and/or RefSeq:
GO:0005515 [from] [protein binding] IEA
GO:0005634 [from] [nucleus] IEA
GO:0005737 [from] [cytoplasm] IEA

InterPro
       IPR002219 Protein kinase C, phorbol ester/diacylglycerol binding
       IPR001781 LIM, zinc-binding
       IPR001841 Zinc finger, RING-type
       IPR001965 Zinc finger, PHD-type

Protein Family ENSF00000000617 : **PHD FINGER**

Query location   : gi|25908280|gb|BU415609.1|BU415609   113 to   510 (-)
Database location : ENSGALT00000016607   976 to   1373 (+)
Genomic location   : 4   35725696 to 35729046 (-)

Alignment score   : 398
E-value   : 3.0e-266
Alignment length   : 398
Percentage identity: 100.00

```
Query:  510 CAGTGTTCAGTGAAGAACTGCAGAACAGCCTTTCACGTCACTTGCGCGTTTGACCGTGGC 451
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  976 CAGTGTTCAGTGAAGAACTGCAGAACAGCCTTTCACGTCACTTGCGCGTTTGACCGTGGC 1035

Query:  450 TTGGAGATGAAGACCATACTGGCAGAGAACGATGAGGTGAAATTTAAGTCATACTGTCCC 391
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1036 TTGGAGATGAAGACCATACTGGCAGAGAACGATGAGGTGAAATTTAAGTCATACTGTCCC 1095

Query:  390 AAGCACAGCTCCACCAAGAGAGCAGATGATGAGACCTTCAGTGACCACCTGTGTCAAGAG 331
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1096 AAGCACAGCTCCACCAAGAGAGCAGATGATGAGACCTTCAGTGACCACCTGTGTCAAGAG 1155

Query:  330 AATGGGAATGGGATTCAGGACAGCTCTCTTCCTGCCCACATTGACCCTTTCCACAGCATG 271
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1156 AATGGGAATGGGATTCAGGACAGCTCTCTTCCTGCCCACATTGACCCTTTCCACAGCATG 1215

Query:  270 GATCAAAACCAGGAGGAAGCCCACAGAGTCAGTCTTCGCAAGCAGAAGCTCCAGCAGCTG 211
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1216 GATCAAAACCAGGAGGAAGCCCACAGAGTCAGTCTTCGCAAGCAGAAGCTCCAGCAGCTG 1275

Query:  210 GAGGATGAATTCTATACCTTTGTCGAGTCTCTGGAAGTGGCTAAAGCACTGCAGCTGCCT 151
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1276 GAGGATGAATTCTATACCTTTGTCGAGTCTCTGGAAGTGGCTAAAGCACTGCAGCTGCCT 1335

Query:  150 GAGGAGCTGGTGGGATTCCTTTATCAGTACTGGAAGCT 113
            |||||||||||||||||||||||||||||||||||||
Sbjct: 1336 GAGGAGCTGGTGGGATTCCTTTATCAGTACTGGAAGCT 1373
```

FUNCTION: Transcriptional coactivator which seems to act by promoting acetylation of nucleosomal histone H4 by HTATIP. Promotes apoptosis. May act as a renal tumor suppressor.
SUBUNIT: Isoform 3 interacts with VHL and HTATIP.
SUBCELLULAR LOCATION: Cytoplasm. Nucleus.
ALTERNATIVE PRODUCTS: 3 named isoforms [FASTA] produced by alternative splicing. Name 1
Synonyms JADE1L
Isoform ID Q6IE81-1
This is the isoform sequence displayed in this entry.

Name 2
Isoform ID Q6IE81-2
Note: No experimental confirmation available.
Features which should be applied to build the isoform sequence: VSP_021045.
Name 3
Synonyms JADE1S
Isoform ID Q6IE81-3
Features which should be applied to build the isoform sequence: VSP_021046, VSP_021047.
TISSUE SPECIFICITY: Highly expressed in kidney. Also present in pancreas, liver and heart (at protein level). Down-regulated in renal cancer cells.
DOMAIN: The 2 PHD-type zinc fingers are required for transcriptional activity.
SIMILARITY: Belongs to the JADE family.
SIMILARITY: Contains 2 PHD-type zinc fingers.

Appendix 5.2 Ontologizer Results (in electronic form - CD)

Appendix 5.3 Genes <=30 FDR in Chicken Pathways

Pathway Group	Pathway Name	Genes
Amino Acid Metabolism		
	Alanine and aspartate metabolism	DARS, AARS
	Cysteine metabolism	SULT4A1, LDHA
	Lysine biosynthesis	AASS
	Lysine degradation	OGDHL, LOC426314, PLOD2, AASS
	Tryptophan metabolism	OGDHL, MYLIP
	Tyrosine metabolism	ADH5
	Urea cycle and metabolism of amino groups	SMS
Carbohydrate Metabolism		
	Citrate cycle (TCA cycle)	SDHA, OGDHL
	Fructose and mannose metabolism	LOC418170
	Galactose metabolism	LOC418170
	Glycolysis / Gluconeogenesis	PGK1, GAPDH, ADH5, LDHA
	Inositol phosphate metabolism	PIP5K1C, ITPKB
	Pentose and glucuronate interconversions	LOC418170
	Pentose phosphate pathway	EPS8L2
	Propanoate metabolism	LDHA
	Pyruvate metabolism	LDHA, LOC418170
	Starch and sucrose metabolism	DDX19B
Energy Metabolism		
	Carbon fixation	PGK1
	Methane metabolism	ADH5
	Oxidative phosphorylation	SDHA, NDUFS1, NDUFA11
Folding, Sorting and Degradation		
	Proteasome	PSMB1
	SNARE interactions in vesicular transport	LOC420134, TSNARE1
Glycan Biosynthesis and Metabolism		
	Glycan structures - biosynthesis 1	HS3ST5, MGAT4A
	Glycosphingolipid biosynthesis – ganglioseries	SLC33A1
	Heparan sulfate biosynthesis	HS3ST5
	N-Glycan biosynthesis	MGAT4A
Lipid Metabolism		
	alpha-Linolenic acid metabolism	ACOX1
	Bile acid biosynthesis	ADH5
	Fatty acid metabolism	ACOX1, ADH5
	Glycerolipid metabolism	ADH5, LOC418170
	Polyunsaturated fatty acid biosynthesis	ACOX1
	Sphingolipid metabolism	SGPL1
Metabolism of Cofactors and Vitamins		
	Biotin metabolism	HLCS
	Folate biosynthesis	DDX19B
	One carbon pool by folate	MTHFS
	Ubiquinone biosynthesis	COQ5

Metabolism of Other Amino Acids		
	beta-Alanine metabolism	SMS
Nucleotide Metabolism		
	Purine metabolism	NME2
	Pyrimidine metabolism	NME2
Signal Transduction		
	Phosphatidylinositol signaling system	PIP5K1C, ITPKB, ITPR2
Translation		
	Aminoacyl-tRNA biosynthesis	AARS, DARS
	Ribosome	RPL10A, RPL4, RPS10, RPLP1, RPS15A
Xenobiotics Biodegradation and Metabolism		
	1- and 2-Methylnaphthalene degradation	ADH5
	3-Chloroacrylic acid degradation	ADH5
	Metabolism of xenobiotics by cytochrome P450	ADH5

Appendix 5.4 Microarray transcripts encoding for (EC:1.6.5.3 1.6.99.3)

Transcripts found on the microarray encoding for the same enzymes (EC:1.6.5.3 1.6.99.3) as LOC770879 (functional candidate gene located under the QTL but not on the microarray)

UniGene	*GeneID*	*t-value*	*p-value*	*Chromosome*
Gga.1735	NDUFA11	-3.59968	0.00465	28
Gga.3251	NDUFS1	-3.29852	7.76E-03	7
Gga.3251	NDUFS1	-2.38217	3.79E-02	7
Gga.3251	NDUFS1	-1.88753	8.77E-02	7
Gga.1524	NDUFB5	-1.8798	8.88E-02	9
Gga.1524	NDUFB5	-1.82485	9.72E-02	9
Gga.7183	NDUFA12	-1.29608	0.223322	1
Gga.1626	NDUFA9	-0.98797	0.345878	1
Gga.3251	NDUFS1	-0.93464	0.371433	7
Gga.22180	NDUFS4	-0.62149	0.547824	Z
Gga.1735	NDUFA11	-0.45931	0.655584	28
Gga.4526	NDUFA4	5.93E-02	0.953828	2
Gga.11296	NDUFB10	7.61E-02	0.940831	14
Gga.42010	RCJMB04_33n14	0.157082	0.878226	NULL
Gga.8284	NDUFS6	0.296055	0.77309	2
Gga.39013	NDUFB4	0.991896	0.344046	1
Gga.4526	NDUFA4	1.080412	0.304681	2
Gga.8285	NDUFB6	2.090865	6.23E-02	14

Chapter 6 General Discussion

The thesis studied the application of targeted genetical genomics to obtain a better characterization of a known QTL through bioinformatics procedures. The bioinformatic characterization of the QTL proved to be a long and challenging process, during which several limitations and problems were encountered. Nevertheless, it was possible to identify genes and pathways that might explain the QTL. The candidate genes and pathways identification was not achieved by a single linear process, but by the repeated application of various interacting procedures, where the outcome of one method would be integrated with, or become the input for, further analysis.

The following sections present key aspects and limitations faced during the process of analysing, interpreting and integrating microarrays with diverse technologies for the dissection of complex traits. In addition, a perspective of the bioinformatics procedures and the new technologies developed for analysing genome genetics are briefly discussed.

6.1 Genetical Genomics: Key aspects, Issues, Limitations and Solutions

In the introduction the concepts of complex traits and genetical genomics were briefly discussed. The dissection of complex traits is highly complex and a multidisciplinary approach is needed for its analysis. 'No gene acts on its own'; genetic changes behave in a chain reaction, complicating the comprehension of the genetics behind complex traits and the downstream changes that they reflect. The typical approaches mostly focus on identifying

the main gene affecting a certain trait or very occasionally try to identify a handful of genes. Nevertheless, identifying one or two genes is not enough. The field attempts to move to a bigger picture (e.g. modelling how changes/perturbations in one gene will affect the expression of other genes and what will be the effect of those changes on the phenotypes). Genetical genomics has been posed as '*the*' promising approach to guide this quest: it exploits the use of gene expression profiles in segregating populations as phenotypes for linkage or association analyses. Genetical genomics makes the assumption that what is causing variation in the traits will be also reflected as variation at the gene expression level. Several studies demonstrated that not only can gene expression differences reflect phenotypic changes but also that gene expression levels can be highly heritable, albeit dependent on tissue, cell types, age, sex and environmental conditions of the samples. These latter factors can significantly affect the gene expression levels by introducing unwanted variations and can lead to the identification of false positives and miss-interpretation of the results (Petretto *et al.* 2006; Ness 2007; Gibson 2008). Hence, the expression levels can be considered to reflect the changes and because of this the experiments have to be carefully planned and the protocols should be strictly followed (e.g. selecting the tissues, ages, environment, etc.).

Main problems faced in genetical genomics:

Since microarrays have so far been the resource to measure the phenotypes in genetical genomics, the discipline also inherits the challenges that come with them. These include the need to correct for and/or control technical variation and find appropriate normalization procedures, as well as the sheer volume of data. A further challenge arises for those species whose genomes are currently being sequenced or only recently completed, for which the

annotations are highly limited and/or changes in the annotations happen constantly. Other important limitations of genetical genomics are the elevated total cost of phenotyping each individual of the study population and the availability of suitable biological samples. These issues and the way they could be handled are discussed further in the following section.

Guidance for normalization and poor quality microarrays:

A helpful approach to assess the behaviour and quality of the microarrays was the use of chip and print-tip spatial plots, box plots and MA plots. The analysis of these plots highlights whole microarrays or regions that might be performing poorly where the gene expression values are highly variable. The statistical correction procedures to be applied to the microarrays can be also guided by these plots, by observing the chip print-tips behaviour.

High-throughput data handling solution:

The solution to the problem of data handling from multiple sources, in this case, was to develop a framework and centralize the information into a relational database (Chapter 4). The identification and selection of tools was based on testing each tool with real data and assessing the 'usefulness' of the results. The methods to integrate the information from various sources and the analyses followed in the microarray studies are very prone to accumulated errors. These errors can originate in many different places, and because of the long process of the analyses and the amount of information handled, is very difficult to identify their origin. As a consequence, in the worst case scenario one mistake can require the complete analysis to be restarted. Therefore, extreme caution is needed in each step of the whole procedure. The use of the framework increased the annotations related to the microarray probes and unified various types of identifiers, allowing

better functional analyses. A similar framework was presented by McCarthy (2007), although their pipeline focused only on GO annotations and analysis.

Although the relational database has the advantage of keeping track of the original gene identifiers and their annotations, it does not resolve the issue of storing the information statically. On the other hand, by knowing how the annotations were obtained and how they are processed, one can re-run the process (framework scripts) and obtain an updated annotation of each identifier provided. In addition, a common 'conflict' with high-throughput data analyses is the loss of the original identifiers and the impossibility of tracking back the source of the annotations, or even tracking back the probe identifiers which the analyses refers to. For example, when using a large input of identifiers to perform KEGG pathway analyses, this program automatically assigns the enzyme identifier from the probes queried, therefore loses the probe identifiers. Furthermore, in many cases we can find several probes in the array coding for the same enzyme, and it should be possible to refer back to the original probe of the microarray. Data centralization solved the problem and allowed to follow the exact transcripts and their associated gene expressions.

Another way of handling the information is to reduce the number of candidate genes by selecting those falling below a certain threshold (i.e. p-values, FDR). This approach is also known as gene-wise analysis. However, the analysis made on a subset of differentially expressed genes restrains a full functional picture of the gene expression variability (e.g changes within pathways, genomic locations, metabolisms). In addition, it is greatly affected by the statistical correction procedures applied to the data. An alternative approach is to investigate the results through predefined gene-sets (Subramanian *et al.* 2005). This method has been widely discussed and

deeply investigated. GSEA represents a more biological-based interpretation of the complete list of ranked genes according to the expression scores through predefined functional gene-sets of pathways, signatures and/or genomic locations (Goeman & Buhlmann 2007; Efron & Tibshirani 2007; Huang *et al.* 2009). However, a great disadvantage at the present, especially for livestock species, is that this method requires prior knowledge and annotations associated to the genes. This analysis was also performed but like the gene ontology analysis, the results were not informative due the poor and unreliable annotations in chicken.

Chen *et al.,*(2008) applied a different approach to this problem. They studied changes through perturbed molecular gene networks of a previously identified obesity QTL, observing how the trait modifies the transcriptional expression networks in trait-relevant tissues. They found that the expression of many transcripts in the QTL region correlate to metabolic traits related to obesity. The modular network approach was also applied to human data to identify genes involved in obesity. Biometric measurements of the individuals were used, and gene expression profiling from blood and fat tissues was performed. They identified overlapping *cis*-acting effects between the tissues (Emilsson *et al.* 2008). Interestingly, the modular networks constructed from the studies made by Emilsson *et al.,* (2008) and Chen *et al.,* (2008) had common characteristics leading to the creation of a 'core' modular network (MEMN) associated with obesity-related traits (Emilsson *et al.* 2008). Yet another interesting example used comparative gene network analysis and biological process genes linked to autism, leading to the identification of new genes interacting with known autism-related genes (Wall *et al.* 2009).

Genetical genomics costs limitations:

As mentioned earlier, one major disadvantage of genetical genomics is the cost of the studies. To reduce the cost of this study it was narrowed down according to the research interests. The presented targeted genetical genomics study concentrated only on a particular QTL of interest, instead of attempting the mapping of genome-wide eQTLs. In addition, the analysis of the expression levels of a known QTL gives more confidence of the link between the genotype and a phenotype. However, the design of the study could be further improved by including more variables (i.e. different tissues, number of samples, samples at various development stages), and also by applying a different microarray design rather than a 'dye-swap' design. The microarray design of the targeted study was limited by the number of biological samples available. Nevertheless, even with such a small experiment, we could get an insight into what could be affecting the trait.

There are very high expectations from the field of genetical genomics, and the reduction of the cost of microarray profiling will allow larger experiments which could reach more statistical power. But dealing with such dimensions of information will still represent a challenge, and further understanding and exploitation of the field could be achieve by the integration of diverse technologies.

Bioinformatics tools and methods:

High-throughput technologies keep on developing and information keeps on growing ever faster, but the methods to analyse it and process it in efficient ways are not developing at the same pace. Visualizing the increasing amount of information comprehensibly is incredibly difficult to achieve.

Constantly in the bioinformatics field there is a particular interest in very specific questions for which tools have not been developed yet (e.g. linking directly the markers between physical and linkage maps), and as soon as one makes the effort and develops the procedures the 'needed' tool becomes available. As a consequence, one has the dilemma of investing time developing the procedures or keep on searching (sometimes waiting) for the procedures to be developed and available. Some advantages of developing the tools yourself are that they will be personalized, so that the resources are 'trusted' and better understanding of the algorithms behind them is gained. A great disadvantage is the amount of time required to develop the tools and debug them. However, the continuous development of new bioinformatics tools and emerging databases creates an overload of repetitive and redundant approaches. This complicates the analyses even further. As discussed in Chapter 2, it is very difficult to consider a single satisfactory tool that is applicable for each step of the analysis process. When dealing with very poorly annotated species better and more reliable results can be expected from the tools/frameworks that use directly the 'base' or primary resources (e.g. NCBI, KEGG).

Validation of candidate genes:

The assumed fundamentals, the prominent exposure to errors and the results obtained from electronically inferred data, create the need for further validation of the derived hypotheses produced by this methodology. However, experimental validation of high-throughput hypotheses becomes almost impossible due the number of experiments that would be required to verify the results of the global analyses. An alternative is to increase the confidence level of the results by investigating the candidate genes and their interactions through '-mining' approaches.

In the thesis four positional candidate genes were identified. The results of the in-depth analyses performed on these candidate genes hypothesized an alternative splicing event on AADAT varying between the two genotypes (Chapter 5). RTPCR analysis was subsequently performed across AADAT (analysis performed by Ian Dunn and colleagues). The RNA utilized on the experiment was obtained from the same animals of the targeted study and the relative density bands produced were quantified. The results confirmed the presence of an alternative splicing event of the exons between exon 2 and 8 varying between the two genotypes.

6.2 Final Remarks

In the recent years, the dissection of complex traits has been studied through genetical genomics and integrative methodologies. The advances in high-throughput technologies and their integrative analytical procedures are opening the way to behavioural, evolutionary and plasticity studies triggered by environmental conditions (Li *et al.* 2006; Landry *et al.* 2006; Gibson 2008). Promising approaches are being developed to investigate further genome-wide gene specificity and regulation through alternative splicing and microRNAs (Yin *et al.* 2008; Wang *et al.* 2008).

Furthermore, the application of next generation sequencing (NGS) technologies to the genomics field could change the designs dramatically and improve significantly the eQTL studies. Next generation sequencing enables the analysis of complete genomes without the need of a pre-known sequenced genome, although it does require a close reference genome to be compared with. This technology opens a great opportunity to those organisms which genome has not been sequenced. And, unlike microarrays,

it does not have cope with sequence cross-hybridization problems nor with microarray experimental designs. In addition, the application of NGS also allows a direct link between genotypes and phenotypes (Shendure & Ji 2008).

A great challenge remains for bioinformatics to handle NGS data (Mardis 2008; Pettersson *et al.* 2009). Pop and Salzberg (2008) investigated some of the challenges that bioinformatics will face, exposing two examples where current procedures had to be adapted according to the needs of new sequencing technologies. One of these examples is the development of MEGAN, a tool created to perform phylogeny analyses (Huson *et al.* 2007); and a second procedure presented was an adaptation of the BLAST algorithm able to mange short sequences produced by this technology (Krause *et al.* 2006). The current bioinformatic procedures may not need to be re-invented, as the process and frameworks can be adapted. For example, the procedures presented here could follow the same 'steps' and processes as they were defined but the tools behind them and their algorithms should be adapted for the new technologies. Furthermore, as the information keeps on growing '-mining' approaches will become more important for research. Literature and data mining are still in their infancy, but great efforts to improve the methodologies and promote their application are undergoing (Altman *et al.* 2008; Shatkay *et al.* 2008; Krallinger *et al.* 2008).

There is a whole new field emerging (bioimaging) which, in its way, will create tools and procedures that will be very useful for any biological based studies. These tools are created with the purpose of processing images, databases and visualization techniques of biological data (Peng 2008).

A great advantage in livestock species and other non-model organisms is that model species studies opened the route to the necessary bioinformatics

tools and procedures. The procedures that have been developed for other species can be easily adapted to livestock needs. The 'ideal' would be to have available bioinformatics tools which allow straight-forward comparisons across species. This would reduce the amount of accumulated errors caused by the various 'processes' (e.g. finding the right identifiers, homologies, annotations and genomic coordinates).

There is no recipe for the dissection of complex traits and neither for the selection of bioinformatics procedures. Clearly, bioinformatics plays a crucial role managing, creating and analysing the data, but ultimately the characterization and interpretation of complex traits is better obtained through integrating various methodologies. The extent and complexity of such studies makes it impossible to perform these analyses by single individuals. I described a framework and annotation procedures that can guide researchers through the complete process and expand the characterization of a known QTL responsible for a certain complex trait. It is highly recommended that researchers should be aware of the various tools and methodologies that lead to closer functional interpretations before performing further experiments for validation. More accurate in-depth analyses performed *in-silico*, could reduce considerably the amount of *in-vivo* experiments needed for the validation of results.

Sometimes, it might be helpful to disassociate ourselves from what is expected to be and how it should be done, in order to obtain the 'answers' we are looking for. Being careful of not falling in the means and losing the goal, in my opinion we need to apply re-engineering and re-structuring in some bioinformatics methods. We get lost in our own data, now the problem is how to analyse it and make it meaningful.

Glossary

Accession number: A unique identifier that is assigned to a single database entry for a DNA or protein sequence.

Algorithm: A systematic procedure for solving a problem in a finite number of steps, typically involving a repetition of operations. Once specified, an algorithm can be written in a computer language and run as a program.

Alignment: Refers to the procedure of comparing two or more sequences by looking for a series of individual characters or character patterns that are in the same order in the sequences. Of the two types of alignment, local and global, a local alignment is generally the most useful.

Allele: One of several alternative forms of a gene occupying a given locus on a chromosome.

Annotation: The prediction of genes in a genome, including the location of protein-encoding genes, the sequence of the encoded proteins, any significant matches to other proteins of known function, and the location of RNA-encoding genes. Predictions are based on the gene models.

Bidirectional replication: Accomplished when two replication forks move away from the same origin in different directions.

Bioinformatics: An interdisciplinary field involving biology, computer science, mathematics, and statistics to analyze biological sequence data, genome content, and arrangement, and to predict the function and the structure of macromolecules.

cDNA: Single-stranded DNA complementary to an RNA, synthesized from it by reverse transcription in-vitro.

cis-acting (eQTL): When the *e*QTL is located in the same region as the gene that is affected.

cis-acting locus: Affects the activity only of DNA sequences on its own molecule of DNA; this property usually implies that the locus does not code for a protein.

Clone: Describes a large number of cells or molecules identical with a single ancestral cell or molecule.

Cluster Analysis: A method for grouping together a set of objects that are most similar from a larger group of related objects. The relationships are based on some criterion of similarity or difference. For sequences, a similarity or distance score or a statistical evaluation of those scores is used.

Codon: A triplet of nucleotides that represents an amino acid or a termination signal.

COG: Clusters of orthologous groups in a set of groups of related sequences in microorganisms and yeast (*S. cervisiae*). These groups are found by whole proteome comparisons and include Orthologs and Paralogs.

Comparative genomics: A comparison of gene numbers, gene locations, and biological functions of genes in the genomes of diverse organism, one objective being to identify groups of genes that play a unique biological role in a particular organism.

Consensus: A single sequence that represents, at each subsequent position, the variation found within corresponding columns of a multiple sequence alignment.

Consensus sequence: Idealized sequence in which each position represents the base most often found when many actual sequences are compared.

Contig: A set of clones that can be assembled into a linear order.

Database: A computerized storehouse of data that provides a standardized way for locating, adding, removing, and changing data.

Deletions: Generated by removal of a sequence of DNA, the regions on either side being joined together.

Dendogram: A form of a tree that lists the compared objects (e.g. sequences or genes in a microarray analysis) in a vertical order and joins related ones by levels of branches extending to one side of the list.

DNA polymerase: Enzyme that synthesizes a daughter strand(s) of DNA (under direction from a DNA template). May be involved in repair or replication.

Domain (protein): Discrete continuous part of the amino acid sequence that can be equated with a particular function.

Enhancer element: A *cis*-acting sequence that increases the utilization of (some) eukaryotic promoters, and can function in either orientation and in any location (upstream or downstream) relative to the promoter.

Epigenetic: Changes influence the phenotype without altering the genotype. Consist of changes in the properties of a cell that are inherited but that do not represent a change in genetic information.

Epistasis: Describes a situation in which expression of one gene obscures the phenotypic effects of another gene.

eQTL: A locus in which genetic variation is associated with the expression variation.

Exon: Any segment of an interrupted gene that is represented in the mature RNA product.

Expected Value (E): In a database similarity search, the probability that an alignment score as good as the one found between a query sequence and a database sequence would be found in as many comparisons between random sequences as was done to find the matching sequence.

False Discovery Rate (FDR): Proportion of false-positive test results among all significant tests.

False negative: A negative data point collected in a data set that was incorrectly reported due to a failure of the test in avoiding negative results.

False positive: A positive data point collected in a data set that was incorrectly reported due to a failure of the test. If the test had correctly measured the data point, the data would have been recorded as negative.

Format (file): Different programs require that information be specified to them in a formal manner, using particular keywords and ordering. This specification is a file format.

Functional genomics: Assessment of the function of genes indentified by between-genome comparisons. The function of a newly identified gene is tested by introducing mutations into the gene and the examining the resultant mutant organism for an altered phenotype.

Gene family: Consists of a set of genes whose exons are related; the members were derived by duplication and variation from some ancestral gene.

Genetical genomics: Process that uses gene expression profiling and marker-based fingerprinting of each individual in a segregating population to analyse factors that underlie variation in gene expression.

Genome: The genetic material of an organism, contained in one haploid set of chromosomes.

Global alignment: Attempts to match as many characters as possible, from end to end, in a set of two or more sequences.

Homolog: A similar component in two organisms (e.g. genes with strongly similar sequences) that can be attributed to a common ancestor of the two organisms during evolution.

Hybridization: Pairing of complementary RNA and DNA strands to give an RNA-DNA hybrid.

Intron: Segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it.

Local alignment: Attempts to align regions of sequences with the highest density of matches. In doing so, one or more islands of subalignments are created in the aligned sequences.

Locus: Position on a chromosome at which the gene for a particular trait resides; locus may be occupied by any one of the alleles for the gene.

Map distance: Measured in cM (centiMorgans) = percent of recombination (sometimes subject to adjustments).

Modification (DNA or RNA): Includes all changes made to the nucleotides after their initial incorporation into the polynucleotide chain.

Mutation: Describes any change in the sequence of genomic DNA.

Normal distribution: The distribution found for many types of data such as body weight, size, and exam scores. The distribution is a bell-shaped curve that is described by a mean and standard deviation of the mean. Local sequence alignment scores between unrelated or random sequences do not follow this distribution but instead the extreme value distribution which has a much extended tail for higher scores.

Oncogenes: Genes whose products have the ability to transform eukaryotic cells so that they grow in a manner analogous to tumour cells.

Open reading frame (ORF): Contains a series of triplets coding for amino acids without any termination codons; sequence is (potentially) translatable into protein.

Orthologs: homologous genes derived through speciation (or vertical descent)

Ontologies: Controlled vocabularies of defined concepts and the relationships between these concepts.

Paralogs: Genes that are related through gene duplication events. These events may lead to the production of a family of related proteins with similar biological functions within a species. Paralogous gene families within a species are identified by using an individual protein as a query in a database similarity search of the entire proteome of an organism. The process is repeated for the entire proteome and the resulting sets of related proteins are then searched for clusters that are most likely to have a conserved domain structure and should represent a paralogous gene family.

Pearson correlation coefficient: A measure of the correlation between two variables that reflects the degree to which the two variables are related. For example, the coefficient is used as a measure of similarity of gene expression in a microarray experiment.

Percent identity: The percentage of the columns in an alignment of two sequences that includes identical amino acids. Columns in the alignment that include gaps are not scored in the calculation.

Percent similarity: The percentage of the columns in an alignment of two sequences that includes either identical amino acids or amino acids that are frequently found substituted for each other in sequences of related proteins (conservative substitutions). These substitutions may be found in an amino acid substitution matrix such as the Dayhoff PAM and Henikoff BLOSUM matrices. Columns in the alignments that include gaps are not scored in the calculation.

Proteome: The entire collection of proteins that are encoded by the genome of an organism. Initially the proteome is estimated by gene prediction and annotation methods but eventually will be revised as more information on the sequence of the expressed genes is obtained.

Pseudogenes: Are inactive but stable components of the genome derived by mutation of an ancestral active gene.

Quantitative trait loci (QTL): Genetic loci or chromosomal regions that contribute to variability in complex quantitative traits.

Regulatory gene: Codes for an RNA or protein product whose function is to control the expression of other genes.

Relational database: Organizes information into tables where each column represents the fields of information that can be stored in a single record. Each row in the table corresponds to a single record. A single database can have many tables and a query language is used to access the data.

Reverse translation: Technique for isolating genes (or mRNAs) by their ability to hybridize with a short oligonucleotide sequence prepared by predicting the nucleic acid sequence from the known protein sequence.

Similarity score (sequence alignment): The sum of the number of identical matches and conservative (high scoring) substitutions in a sequence alignment divided by the total number of aligned sequence characters. Gaps are usually ignored.

Splice sites: Are the sequences immediately surrounding the exon-intron boundaries.

Splicing: Describes the removal of introns and joining of exons in RNA; thus introns are spliced out, while exons are spliced together.

Stop codons: Are the three triplets (UAA, UAG, UGA) which terminate protein synthesis.

Synteny : The presence of a set of homologous genes in the same order on two genomes.

trans: Configuration of two sites refers to their presence on two different molecules of DNA (chromosomes).

Transcription: Is the synthesis of RNA on a DNA template.

Translation: Is the synthesis of protein on the mRNA template.

Variegation: Of phenotype is produced by a change in genotype during somatic development.

Zinc finger protein: Has a repeated motif of amino acids with characteristic spacing of cysteines that may be involved in binding zinc; is characteristic of some proteins that bind DNA and/or RNA.

Literature Cited

2004 EDITORIAL. *Nucleic Acids Research* **32**: D1.

Allison, D. B., X. Cui, G. P. Page, and M. Sabripour, 2006 Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**: 55-65.

Altman, R., C. Bergman, J. Blake, C. Blaschke, A. Cohen *et al.* 2008 Text mining for biology - the way forward: opinions from leading scientists. *Genome Biology* **9**: S7.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.

Andersen, M. C., P. G. Engstrom, S. Lithwick, D. Arenillas, P. Eriksson *et al.* 2008 In Silico Detection of Sequence Variations Modifying Transcriptional Regulation. *PLoS Comput Biol* **4**: e5.

Andersson, L., 2001 Genetic dissection of phenotypic diversity in farm animals. *Nature Reviews Genetics* **2**: 130-138.

Arbilly, A. P., 2006 An integrative approach for the identification of quantitative trait loci. *Animal Genetics* **37**: 7-9.

Asano, H., T. Murate, T. Naoe, H. Saito, and G. Stamatoyannopoulos, 2004 Molecular cloning and characterization of ZFF29: a protein containing a unique Cys(2)His(2) zinc-finger motif. *Biochemical Journal* **384**: 647-653.

Baird, D., P. Johnstone, and T. Wilson, 2004 Normalization of microarray data using a spatial mixed model analysis which includes splines. *Bioinformatics* **20**: 3196-3205.

Bauer, S., S. Grossmann, M. Vingron, and P. N. Robinson, 2008 Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24**: 1650-1651.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**: 289-300.

Borevitz, J. O., and J. Chory, 2004 Genomics tools for QTL analysis and gene discovery. *Current Opinion in Plant Biology* **7**: 132-136.

Brazma, A., and J. Vilo, 2000 Gene expression data analysis. *FEBS Letters* **480**: 17-24.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**: 752-755.

Brown, P. O., and D. Botstein, Exploring the new world of the genome with DNA microarrays. *Nat Genet* .

Bueno Filho, J. S., S. G. Gilmour, and G. J. Rosa, 2006 Design of Microarray Experiments for Genetical Genomics Studies. *Genetics* genetics.

Burnside, J., P. Neiman, J. Tang, R. Basom, R. Talbot *et al.* 2005 Development of a cDNA array for chicken gene expression analysis. *BMC Genomics* **6**.

Burt, D. W., 2005 Chicken genome: Current status and future opportunities. *Genome Research* **15**: 1692-1698.

Burt, D. W., and S. J. White, 2007 Avian genomics in the 21st century. *Cytogenetic and genome research* **117**: 6-13.

Butte, A., 2002 The use and analysis of microarray data. *Nature Reviews Drug Discovery* **1**: 951-960.

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher *et al.* 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37**: 225-232.

Cavalieri, D., and C. De Filippo, 2005 Bioinformatic methods for integrating whole-genome expression results into cellular networks. *Drug Discovery Today* **10**: 727-734.

Chen, Y., J. Zhu, P. Y. Lum, X. Yang, S. Pinto *et al.* 2008 Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**: 429-435.

Chesler, E. J., L. Lu, S. Shou, Y. Qu, J. Gu *et al.* 2005 Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* **37**: 233-242.

Clop, A., F. Marcq, H. Takeda, D. Pirottin, X. Tordoir *et al.* 2006 A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet* **advanced online publication**.

Cogburn, L. A., T. E. Porter, M. J. Duclos, J. Simon, S. C. Burgess *et al.* 2007 Functional Genomics of the Chicken A Model Organism. *Poultry Science* **86**: 2059-2094.

Cui, X., I. D. Vivo, R. Slany, A. Miyamoto, R. Firestein *et al.* 1998 Association of SET domain and myotubularin-related proteins modulates growth control. *Nat Genet* **18**: 331-337.

Curtis, R. K., M. Oresic, and A. Vidal-Puig, 2005 Pathways to the analysis of microarray data. *Trends in Biotechnology* **23**: 429-435.

Dahlquist, K. D., N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin, 2002 GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics* **31**: 19-20.

de Koning, D. J., C. P. Cabrera, and C. S. Haley, 2007 Genetical Genomics: Combining Gene Expression with Marker Genotypes in Poultry. *Poultry Science* **86**: 1501-1509.

de Koning, D. J., and C. S. Haley, 2005 Genetical genomics in humans and model organisms. *Trends in Genetics* **21**: 377-381.

Dennis, G., B. Sherman, D. Hosack, J. Yang, W. Gao *et al.* 2003 DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* **4**: R60.

Draghici, S., P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, 2003 Global functional profiling of gene expression. *Genomics* **81**: 98-104.

Drake, T., E. Schadt, and A. Lusis, 2006 Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. *Mammalian Genome* **17**: 466-479.

Drickamer K., and A. J. Fadden, 2002 *Genomic analysis of C-type lectins*.

Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent, 1999 Expression profiling using cDNA microarrays. *Nat Genet* **21**: 10-14.

Efron, B., and R. Tibshirani, 2007 On testing the significance of sets of genes. *Ann.Appl.Stat.* **1**: 107-129.

Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink *et al.* 2008 Genetics of gene expression and its effect on disease. *Nature* **452**: 423-428.

Fatufe, A. A., R. Timmler, and M. Rodehutscord, 2004 Response to lysine intake in composition of body weight gain and efficiency of lysine utilization of growing male chickens from two genotypes. *Poultry Science* **83**: 1314-1324.

Fisher, P., C. Hedeler, K. Wolstencroft, H. Hulme, H. Noyes *et al.* 2007 A systematic strategy for large-scale analysis of genotype phenotype

correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Research* **35**: 5625-5633.

Friedman-Einat, M., T. Boswell, G. Horev, G. Girishvarma, I. C. Dunn *et al.* 1999 The Chicken Leptin Gene: Has It Been Cloned? *General and Comparative Endocrinology* **115**: 354-363.

Fu, H., and H. K. Dooner, 2002 Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 9573-9578.

Fu, J., and R. C. Jansen, 2006 Optimal Design and Analysis of Genetic Studies on Gene Expression. *Genetics* **172**: 1993-1999.

Galperin, M. Y., and G. R. Cochrane, 2009 Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Research* **37**: D1-D4.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling *et al.* 2004 Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**.

Gibson, G., 2008 The environmental contribution to gene expression profiles. *Nature Reviews Genetics* **9**: 575-581.

Gibson, G., and B. Weir, 2005 The quantitative genetics of transcription. *Trends in Genetics* **21**: 616-623.

Glazier, A. M., J. H. Nadeau, and T. J. Aitman, 2002 Finding Genes That Underlie Complex Traits. *Science* **298**: 2345-2349.

Goeman, J. J., and P. Buhlmann, 2007 Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**: 980-987.

Goh, D. L. M., A. Patel, G. H. Thomas, G. S. Salomons, D. S. M. Schor *et al.* 2002 Characterization of the human gene encoding [alpha]-aminoadipate aminotransferase (AADAT). *Molecular Genetics and Metabolism* **76**: 172-180.

Grant, G. R., J. Liu, and C. J. Stoeckert, Jr., 2005 A practical false discovery rate approach to identifying patterns of differential expression in microarray data. *Bioinformatics* **21**: 2684-2690.

Groenen, M. A. M., H. H. Cheng, N. Bumstead, B. F. Benkel, W. E. Briles *et al.* 2000 A Consensus Linkage Map of the Chicken Genome. *Genome Research* **10**: 137-147.

Hoffmann, R., and A. Valencia, 2005 Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* **21**: 252-258.

Hoheisel, J. D., 2006 Microarray technology: beyond transcript profiling and genotype analysis. *Nature Reviews Genetics* **7**: 200-210.

Hu, Z. L., and J. M. Reecy, 2007 Animal QTLdb: beyond a repository. *Mammalian Genome* **18**: 1-4.

Huang, D. W., B. T. Sherman, and R. A. Lempicki, 2009 Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**: 1-13.

Hubbard, T., D. Andrews, M. Caccamo, G. Cameron, Y. Chen *et al.* 2005 Ensembl 2005. *Nucleic Acids Research* **33**: D447-D453.

Hubner, N., C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz *et al.* 2005 Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* **37**: 243-253.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster, 2007 MEGAN analysis of metagenomic data. *Genome Research* **17**: 377-386.

Ihmels, J., G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv *et al.* 2002 Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31**: 370-377.

Jansen, R. C., 2003 Studying complex biological systems using multifactorial perturbation. *Nature Reviews Genetics* **4**: 145-151.

Jansen, R. C., and J. P. Nap, 2001 Genetical genomics: the added value from segregation. *Trends in Genetics* **17**: 388-391.

Jensen, L. J., J. Saric, and P. Bork, 2006 Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics* **7**: 119-129.

Jin, C., H. Lan, A. D. Attie, G. A. Churchill, D. Bulutuglo *et al.* 2004 Selective Phenotyping for Increased Efficiency in Genetic Mapping Studies. *Genetics* **168**: 2285-2293.

Kadarmideen, H., P. von Rohr, and L. Janss, 2006 From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. *Mammalian Genome* **17**: 548-564.

Keller, B., K. Emrich, N. Hoecker, M. Sauer, F. Hochholdinger *et al.* 2005 Designing a microarray experiment to estimate dominance in maize (Zea mays L.). *TAG Theoretical and Applied Genetics* **111**: 57-64.

Kent, W. J., 2002 BLATΓÇöThe BLAST-Like Alignment Tool. *Genome Research* **12**: 656-664.

Kerr, M. K., and G. A. Churchill, 2001 Experimental design for gene expression microarrays. *Biostatistics* **2**: 183-201.

Khatri, P., and S. Draghici, 2005 Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21**: 3587-3595.

Kirst, M., A. A. Myburg, J. P. G. De Leon, M. E. Kirst, J. Scott *et al.* 2004 Coordinated Genetic Regulation of Growth and Lignin Revealed by Quantitative Trait Locus Analysis of cDNA Microarray Data in an Interspecific Backcross of Eucalyptus. *PLANT PHYSIOLOGY* **135**: 2368-2378.

Koehler, J., J. Baumbach, J. Taubert, M. Specht, A. Skusa *et al.* 2006 Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* btl081.

Krallinger, M., A. Valencia, and L. Hirschman, 2008 Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology* **9**: S8.

Krause, L., N. N. Diaz, D. Bartels, R. A. Edwards, A. Puhler *et al.* 2006 Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics* **22**: e281-e289.

Kuo, A. Y., M. A. Cline, E. Werner, P. B. Siegel, and D. M. Denbow, 2005 Leptin effects on food and water intake in lines of chickens selected for high or low body weight. *Physiology & Behavior* **84**: 459-464.

Labib, M., 2003 The investigation and management of obesity. *Journal of Clinical Pathology* **56**: 17-25.

Lam, A. C., J. Fu, R. C. Jansen, C. S. Haley, and D. J. de Koning, 2008 Optimal Design of Genetic Studies of Gene Expression With Two-Color Microarrays in Outbred Crosses. *Genetics* **180**: 1691-1698.

Landry, C. R., J. Oh, D. L. Hartl, and D. Cavalieri, 2006 Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* **366**: 343-351.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan *et al.* 2007 Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

Lee, H., R. Habas, and C. bate-Shen, 2004 Msx1 Cooperates with Histone H1b for Inhibition of Transcription and Myogenesis. *Science* **304**: 1675-1678.

Leung, Y. F., and D. Cavalieri, 2003 Fundamentals of cDNA microarray data analysis. *Trends in Genetics* **19**: 649-659.

Li, Y., R. Breitling, and R. C. Jansen, 2008 Generalizing genetical genomics: getting added value from environmental perturbation. *Trends in Genetics* **24**: 518-524.

Li, Y., O. A. Ivarez, E. W. Gutteling, M. Tijsterman, J. Fu *et al.* 2006 Mapping Determinants of Gene Expression Plasticity by Genetical Genomics in C. elegans. *PLoS Genetics* **2**: e222.

Liu, H. C., H. H. Cheng, V. Tirunagaru, L. Sofer, and J. Burnside, 2001 A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping. *Animal Genetics* **32**: 351-359.

Lockhart, D. J., and E. A. Winzeler, 2000 Genomics, gene expression and DNA arrays. *Nature* **405**: 827-836.

Marchler-Bauer, A., J. B. Anderson, M. K. Derbyshire, C. Weese-Scott, N. R. Gonzales *et al.* 2007 CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Research* **35**: D237-D240.

Mardis, E. R., 2008 The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**: 133-141.

McCarthy, F., 2007 GOing from functional genomics to biological significance. *Cytogenetic and genome research* **117**: 278-287.

McCarthy, F., N. Wang, G. B. Magee, B. Nanduri, M. Lawrence *et al.* 2006 AgBase: a functional genomics resource for agriculture. *BMC Genomics* **7**: 229.

Monks, S. A., A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak *et al.* 2004 Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics* **75**: 1094-1105.

Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens *et al.* 2004 Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743-747.

Mount D. M., 2001 *Bioinformatics: Sequence and Genome Analysis*.

Nagase, T., K. i. Ishikawa, R. Kikuno, M. Hirosawa, N. Nomura *et al.* 1999 Prediction of the Coding Sequences of Unidentified Human Genes.XV. The Complete Sequences of 100 New cDNA Clones from Brain Which Code for Large Proteins in vitro. *DNA Research* **6**: 337-345.

Natale, D., C. Arighi, W. Barker, J. Blake, T. C. Chang *et al.* 2007 Framework for a Protein Ontology. *BMC Bioinformatics* **8**: S1.

Navarro, P., P. M. Visscher, Knott S.A., Burt D.W., P. M. Hocking *et al.* 2005 Mapping of quantitative trait loci affecting organ weights and blood variablesin a broiler layer cross. *Br Poult Sci.* **46**: 430-442.

Ness, S. A., 2007 Microarray analysis: basic strategies for successful experiments. *Mol Biotechnol* **36**: 205.

Ogata, H., S. Goto, W. Fujibuchi, and M. Kanehisa, 1998 Computation with the KEGG pathway database. *Biosystems* **47**: 119-128.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono *et al.* 1999 KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **27**: 29-34.

Olson, N. E., 2006 The Microarray Data Analysis Process: From Raw Data to Biological Significance. *NeuroRX* **3**: 373-383.

Pankaj Jaiswal, Shulamit Avraham, Katica Ilic, E. A. Kellogg, S. McCouch *et al.* 2005 Plant Ontology (PO): a Controlled Vocabulary of Platn Structures and Growth Stages. *Comparative and Functional Genomics* **6**: 388-397.

Parkinson, H., U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino *et al.* 2005 ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research* **33**: D553-D555.

Pastinen, T., B. Ge, and T. J. Hudson, 2006 Influence of human genome polymorphism on gene expression. *Human Molecular Genetics* **15**: R9-16.

Peltonen, L., and V. A. McKusick, 2001 Genomics and Medicine: Dissecting Human Disease in the Postgenomic Era. *Science* **291**: 1224-1229.

Peng, H., 2008 Bioimage informatics: a new area of engineering biology. *Bioinformatics* **24**: 1827-1836.

Petretto, E., J. Mangion, N. J. Dickens, S. A. Cook, M. K. Kumaran *et al.* 2006 Heritability and Tissue Specificity of Expression Quantitative Trait Loci. *PLoS Genetics* **2**: e172.

Pettersson, E., J. Lundeberg, and A. Ahmadian, 2009 Generations of sequencing technologies. *Genomics* **93**: 105-111.

Philippi, S., and J. Kohler, 2006 Addressing the problems with life-science databases for traditional uses and systems biology. *Nature Reviews Genetics* **7**: 482-488.

Ponsuksili, S., E. Murani, K. Schellander, M. Schwerin, and K. Wimmers, 2005 Identification of functional candidate genes for body composition by expression analyses and evidencing impact by association analysis and mapping. *Biochim.Biophys.Acta* **1730**: 31-40.

Pop, M., and S. L. Salzberg, 2008 Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24**: 142-149.

Qiu, X., A. Brooks, L. Klebanov, and A. Yakovlev, 2005 The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics* **6**: 120.

Quackenbush, J., 2001 Computational analysis of microarray data. *Nature Reviews Genetics* **2**: 418-427.

Reimers, M., and J. Weinstein, 2005 Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases. *BMC Bioinformatics* **6**: 166.

Reiner, A., D. Yekutieli, and Y. Benjamini, 2003 Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**: 368-375.

Ron, M., and J. I. Weller, 2007 From QTL to QTN identification in livestock - winning by points rather than knock-out: a review. *Animal Genetics* **38**: 429-439.

Rosa, G. J. M., N. de Leon, and A. J. M. Rosa, 2006 Review of microarray experimental design strategies for genetical genomics studies. *Physiological Genomics* **28**: 15-23.

Rosenbaum, M., and R. L. Leibel, 1998 Leptin: A molecule integrating somatic energy stores, energy expenditure and fertility. *Trends in Endocrinology and Metabolism* **9**: 117-124.

Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards *et al.* 2005 An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37**: 710-717.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che *et al.* 2003 Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297-302.

Seaton, G., C. S. Haley, S. A. Knott, M. Kearsey, and P. M. Visscher, 2002 QTL Express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics* **18**: 339-340.

Secko, D., 2005 Genetics embraces expression. *Scientist* **19**: 26-27.

Sewalem, A., D. M. Morrice, A. Law, D. Windsor, C. S. Haley *et al.* 2002 Mapping of quantitative trait loci for body weight at three, six, and nine weeks of age in a broiler layer cross. *Poultry Science* **81**: 1775-1781.

Sharp, P. J., I. C. Dunn, D. Waddington, and T. Boswell, 2008 Chicken leptin. *General and Comparative Endocrinology* **158**: 2-4.

Shatkay, H., F. Pan, A. Rzhetsky, and W. J. Wilbur, 2008 Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics* **24**: 2086-2093.

Shendure, J., and H. Ji, 2008 Next-generation DNA sequencing. *Nat Biotech* **26**: 1135-1145.

Shiu, S. H., and J. O. Borevitz, 2006 The next generation of microarray research: applications in evolutionary and ecological genomics. *Heredity* **100**: 141-149.

Siegel, P. B., J. B. Dodgson, and L. Andersson, 2006 Progress from Chicken Genetics to the Chicken Genome. *Poultry Science* **85**: 2050-2060.

Smith, B., M. Ashburner, C. Rosse, J. Bard, W. Bug *et al.* 2007 The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech* **25**: 1251-1255.

Smyth, G. K., 2004 Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat.Appl.Genet.Mol.Biol.* **3**: Article3.

Smyth, G. K., and T. Speed, 2003 Normalization of cDNA microarray data. *Methods* **31**: 265-273.

Sonmez, K., N. T. Zaveri, I. A. Kerman, S. Burke, C. R. Neal *et al.* 2009 Evolutionary Sequence Modeling for Discovery of Peptide Hormones. *PLoS Comput Biol* **5**: e1000258.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. *Proc.Natl.Acad.Sci.U.S.A* **100**: 9440-9445.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert *et al.* 2005 Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 15545-15550.

Tavazoie, S., J. D. Hughes, M. Campbell, R. J. Chao, and G. M. Church, 1999 Systematic determination of genetic network architecture. *Nature Genetics* **22**: 281-285.

Tesseraud, S., R. Peresson, J. Lopes, and A. M. Chagneau, 1996 Dietary lysine deficiency greatly affects muscle and liver protein turnover in growing chickens. *British Journal of Nutrition* **75**: 853-865.

Torniainen, S., R. Freddara, T. Routi, C. Gijsbers, C. Catassi *et al.* 2009 Four novel mutations in the lactase gene (LCT) underlying congenital lactase deficiency (CLD). *BMC Gastroenterology* **9**: 8.

Twigger, S. N., M. Shimoyama, S. Bromberg, A. E. Kwitek, H. J. Jacob *et al.* 2007 The Rat Genome Database, update 2007--Easing the path from disease to data and back again. *Nucleic Acids Research* **35**: D658-D662.

Van Laere, A. S., M. Nguyen, M. Braunschweig, C. Nezer, C. Collette *et al.* 2003 A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**: 832-836.

Vastrik, I., P. D'Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath *et al.* 2007 Reactome: a knowledge base of biologic pathways and processes. *Genome Biology* **8**: R39.

Wall, D. P., F. J. Esteban, T. F. DeLuca, M. Huyck, T. Monaghan *et al.* 2009 Comparative analysis of neurological disorders focuses genome-wide search for autism genes. *Genomics* **93**: 120-129.

Wang, D., and D. Nettleton, 2006 Identifying Genes Associated with a Quantitative Trait or Quantitative Trait Locus via Selective Transcriptional Profiling. *Biometrics* **62**: 504-514.

Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang *et al.* 2008 Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470-476.

Wayne, M. L., and L. M. McIntyre, 2002 Combining mapping and arraying: An approach to candidate gene identification. *Proceedings of the National Academy of Sciences* **99**: 14903-14906.

Woloshin, P., K. Song, C. Degnin, A. M. Killary, D. J. Goldhamer *et al.* 1995 MSX1 inhibits MyoD expression in fibroblast x 10T1/2 cell hybrids. *Cell* **82**: 611-620.

Wong, K. M., M. A. Suchard, and J. P. Huelsenbeck, 2008 Alignment Uncertainty and Genomic Analysis. *Science* **319**: 473-476.

Wu, K., Y. Yang, C. Wang, M. A. Davoli, M. D'Amico *et al.* 2003 DACH1 Inhibits Transforming Growth Factor-{beta} Signaling through Binding Smad4. *Journal of Biological Chemistry* **278**: 51673-51684.

Yang Y.H., Dudoiit S., Luu P. & Speed T.P. Normalization for cDNA Microarray Data. 2006.
Ref Type: Generic

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng *et al.* 2002 Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**: e15.

Yang, Y. H., and T. Speed, 2002 Design issues for cDNA microarray experiments. *Nature Reviews Genetics* **3**: 579-588.

Yao, B., S. Rakhade, Q. Li, S. Ahmed, R. Krauss *et al.* 2004 Accuracy of cDNA microarray methods to detect small gene expression changes induced by neuregulin on breast epithelial cells. *BMC Bioinformatics* **5**: 99.

Yin, J. Q., R. C. Zhao, and K. V. Morris, 2008 Profiling microRNA expression with microarrays. *Trends in Biotechnology* **26**: 70-76.

Yvert, G., R. B. Brem, J. Whittle, J. M. Akey, E. Foss *et al.* 2003 Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet* **35**: 57-64.

Zhou, M. I., H. Wang, J. J. Ross, I. Kuzmin, C. Xu *et al.* 2002 The von Hippel-Lindau Tumor Suppressor Stabilizes Novel Plant Homeodomain Protein Jade-1. *Journal of* Biological *Chemistry* **277**: 39887-39898.