Functional Prediction of Genetic Variation within and between Two Chicken Lines Selected for Body-Weight

With Bioinformatic Methods

Xidan Li

Faculty of Veterinary Medicine and Animal Sciences
Department of Clinical Sciences
Uppsala

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2014

Acta Universitatis Agriculturae Sueciae
2014:24

Cover: The picture is made by my dear friend – Dr. Ronald Nelson.
(photo: Matrix Reloaded)

ISSN 1652-6880
ISBN (Print version) 978-91-576-7996-3
ISBN (Electronic version) 978-91-576-7997-0

Print: SLU Service/Repro, Uppsala 2014

Functional Prediction of Genetic Variation within and between Two Chicken Lines Selected for Body-Weight - With Bioinformatic Methods

Abstract
Identifying genetic variation influencing complex traits is often a big challenge. Paul Siegel at the Virginia Polytechnic Institute and State University (USA) initiated a breeding experiment in the 1950s, where White Plymouth Rock (WPR) chicken lines were bi-directionally selected for body-weight at 56 days of age. After more than 50 generations of selection, the High Weight Selected (HWS) line is more than 10-fold bigger than the Low Weight Selected (LWS) line. These HWS and LWS lines have become a good model to investigate the genetic mechanisms underlying the body weight changes under long-term selection. Moreover, as a result of the recently rapid development of next generation sequencing technologies, with high throughput, a large number of genetics polymorphisms have been identified and can be used to explore the genetic factors underlying complex traits. In this thesis, we used NGS resequencing data and several leading databases to search for genes and mutations within previously mapped epistasic QTL regions, which could explain the differences in growth-related traits between the HWS and LWS lines. In consequence, a number of genetic factors have been detected and provide a good basis for further experimental investigation in relation to the observed effects on growth and other metabolic traits. Additionally, we also developed three softwares, which can be useful in the process of identifying genes and variations with phenotypic effects. One of these softwares were also applied within genetic studies in this thesis. Our softwares could be widely applied in many species and are likely to benefit many other research projects.

Keywords: chicken, body weight, QTL, epistasis, NGS, SNP, amino acid substitution, physico-chemical property, protein motif, expression phenotype.

Author's address: Xidan Li, SLU, Department of Clinical Sciences,
P.O. Box 7078, 750 07 Uppsala, Sweden
E-mail: Xidan.Li@slu.se

Dedication

To my beloved parents

谨以此书献给我亲爱的爸爸妈妈，感谢你们对我的无私的支持！

李锡丹

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I **Xidan Li,** Muhammad Ahsan, Andreas E Lundberg, Marcin Kierczak, Paul B Siegel, Örjan Carlborg, Stefan Marklund. Identification of candidate genes and SNPs within epistatic chicken growth QTL regions. (Manuscript)

II Muhammad Ahsan, **Xidan Li,** Andreas E Lundberg, Marcin Kierczak, Paul B Siegel, Örjan Carlborg, Stefan Marklund. (2013) Identification of candidate genes and mutations in QTL regions for chicken growth using bioinformatic analysis of NGS and SNP-chip data. *Frontiers in Genetics, doi: 10.3389/fgene.2013.00226*

III **Xidan Li**§, Marcin Kierczak§, Xia Shen, Muhammad Ahsan, Örjan Carlborg, Stefan Marklund. (2013) PASE: a novel method for functional prediction of amino acid substitutions based on physicochemical properties. *Frontiers in Genetics, doi: 10.3389/fgene.2013.00021*

IV **Xidan Li**, Zheya Sheng and Stefan Marklund. Profat: A novel method for prediction of novel protein function using the motif profile classification prior to sequence alignments. (Manuscript)

V **Xidan Li,** Martin Norling and Ronald Nelson. DIPT: Software for detecting interactions and pathways between expression phenotypes and associated QTL/GWAS defined region. (Manuscript)

§ Authors contributed equally

The contribution of Xidan Li to the papers included in this thesis was as follows:

I Drafted manuscript and carried out the region-targeted computation and analysis using the different sources of data and took part in the planning of the study.

II Took part in the region-targeted computation and analysis using the different sources of data and contributed to the planning of the study.

III Partly planned the study, implemented and evaluated the algorithm and drafted the manuscript.

IV Conceptualized the study, developed the algorithm, compiled the script and drafted the manuscript.

V Compiled the scripts, implemented the algorithm and drafted the manuscript.

Related works by author:

(not included in the thesis)

I **Xidan Li** and Stefan Marklund. S.I.T: a novel software for identifying the candidate SNPs in QTL/GWAS defined regions. (Manuscript in preparation)

II Ronald M Nelson, Mats E Pettersson, **Xidan Li** and Örjan Carlborg. Variance heterogeneity in Saccharomyces cerevisiae expression data: trans-regulation and epistasis. *PLoS ONE 8(11): e79507. doi:10.1371/journal.pone.0079507*

III **Xidan Li**, Xiaoding Liu, Javad Nadaf, Elisabeth Le Bihan-Duval, Ian Dunn, Richard Talbot and Dirk-Jan De Koning. Using targeted re-sequencing for identification of candidate genes and SNPs for a QTL affecting the PH value of chicken meat. (Manuscript)

IV Deniz M. Özata, **Xidan Li**, Stefan Marklund, Patrick Scicluna, Ersen Kavak, Rickard Sandberg, Catharina Larsson and Weng-Onn Lui. Identification and characterization of small RNAs in human testicular germ cell tumors using a deep sequencing approach. (Manuscript)

Abbreviations

QTL	Quantitative Trait Locus
GWA	Genome Wide Association
NGS	Next Generation Sequencing
HWS	High Weight Selected
LWS	Low Weight Selected
SNP	Single Nucleotide Polymorphism
AAS	Amino Acid Substitution
RJF	Red Jungle Fowl
WPR	White Plymouth Rock
MSA	Multiple Sequence Alignment
CPU	Central Processing Unit
GPU	Graphics Processing Unit
FSV	Flanking-SNP-Value
PDB	Protein Data Bank
VEP	Variant Effect Predictor
CDS	Combined Data Score

1 Introduction

Since humans started to domesticate animals, chicken has been selectively bred for thousands of years, which has resulted in a diversity of phenotypes. Historically, chicken has been widely used as a good source for investigating the genetic basis of phenotypic traits. One good example is the two divergent chicken lines for body weight developed by Paul Siegel at the Virginia Polytechnic Institute and State University (USA), where two chicken lines from a single founder population of White Plymouth Rock chickens were divergently selected for body weight at 56 days age. After 50 generations of selection, the High Weight Selected (HWS) line was more than ten-fold bigger than the Low Weight Selected (LWS) at 56 days of age.

A number of metabolic traits to the selection response have shown significant differences between the HWS and LWS lines, for example, a remarkable divergence in appetite, where the HWS lines are hyperphagic whereas the LWS chickens have a reduced appetite including some chickens that never start to eat (Noble *et al*., 1993).

Previous studies have shown that the selection responses are caused by many quantitative trait loci (QTL), where each of them contributes with a small effect (Jacobsson *et al*., 2005, Wahlberg *et al*., 2009; Besnier *et al*. 2011), as well as by strong epistasis (Carlborg *et al*., 2006; Pettersson *et al*., 2011). A great challenge is to explore the genetic architecture underlying the selection response.

With the rapid development of technologies for high throughput sequence analysis like Next Generation Sequencing (NGS) methods, identifications of genetic polymorphisms have been dramatically accelerated. The studies in this thesis include the use of resequencing data and available bioinformatics resources as well as software development for functional prediction to unravel the genetic

mechanisms that contributes to the observed selection responses in the HWS and LWS lines.

2 Background

This thesis identifies the genetic elements and structures responsible for two chicken lines divergently selected for juvenile bodyweight. New methods and software were also developed as described in this thesis, to facilitate the identification of genetic factors underlying the phenotypic differences between HWS and LWS lines.

2.1 The Chicken as an Animal Model

Animals used for breeding have been selected based on their phenotype for 9,000 – 12,000 years (Clutton-Brock, 1995). This approach, therefore, is called selective breeding. However, it is difficult to evaluate the genetic factors influencing long-term artificial selective breeding for a quantitative trait (Dunnington and Siegel 1985). As chicken has experienced the long time selection and resulted the diversity of phenotype, it constitutes a good model to investigate the genetic variances underlying phenotypic traits.

In this thesis, White Plymouth Rock chicken lines divergently selected for body-weight were used as a model to investigate genetic mechanisms underlying growth-related traits (Dunnington and Siegel, 1996, Dunnington *et al*., 2013, Zhao *et al*., 2013).

2.1.1 The Chicken Model for Metabolic Traits – Body Weight

Body weight has been considered as a trait with moderate heritability. It has been widely applied as the benchmark of selection experiments, which explores both direct and correlated genetics changes in artificial selective populations (Dunnington and Siegel 1985).

In one such selection experiment, managed by Paul Siegel since the 1950s, bi-directional selection for body-weight at 56 days of ages was

carried out with White Plymouth Rock chickens (Dunnington and Siegel, 1996). After more than 50 generations of selection, the phenotype of two chicken lines have undergone a radical change, where the HWS line has become more than 10-fold bigger than the LWS lines (Marquez *et al*., 2010, Dunnington *et al*., 2013, Zhao *et al*., 2013).

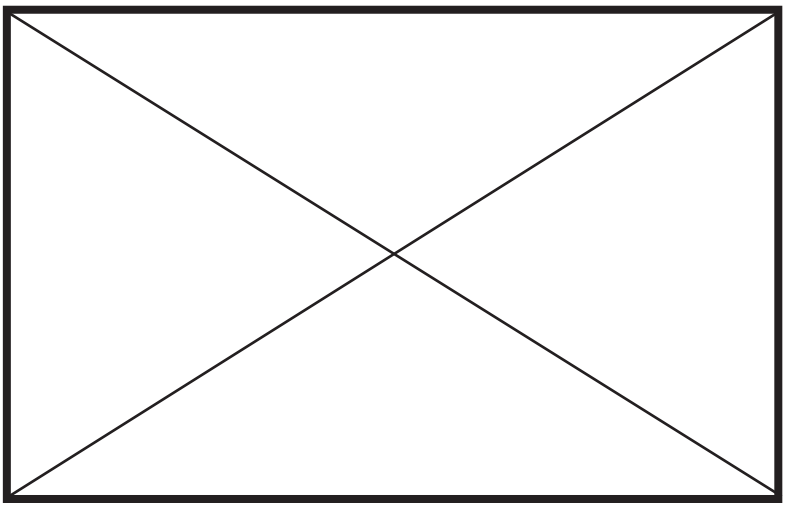

Figure 1. Two extreme chicken lines through artificial selection for body-weight at eight weeks age have been created at Virginia Tech, USA since 1950's. After 50 generations selection, the HWS line (right) has become more than ten-fold bigger than LWS line. (Marquez *et al*., 2010, Dunnington *et al*., 2013, Zhao *et al*., 2013).

Variation between the HWS and LWS lines has been observed in a number of metabolic traits of relevance for the selection responses of body weight, including remarkable differences in appetite, where HWS chickens are hyperphagic whereas LWS chickens have a reduced appetite, including some chickens that never start to eat. This makes these chicken lines an interesting model for e.g. anorexia (Dunnington and Siegel, 1996, Newmyer *et al*., 2010, Newmyer *et al*., 2013, Xu *et al*., 2011). Moreover, the antibody response to SRBC via either intramuscular injections or intravenous injections has shown significant differences between two lines (Boa-Amponsem *et al*., 1998; Paramentier *et al*., 1996; Pinard van der lann *et al*., 1998). The findings demonstrate that the immune system has undergone the negative

response under selection for increased growth, which implies the resources are on competition between growth and immune system (Jacobsson *et al*., 2005).

2.1.2 Mapped QTL Regions with effects on Chicken Body Weight

Previous studies have demonstrated that the selection response in the Virginia chicken lines is caused by many quantitative trait loci (QTL), each of which contributes a small effect (Jacobsson *et al*., 2005, Wahlberg *et al*., 2009), as well as significant epistasis (Carlborg *et al*., 2006). Moreover, interacting loci on chromosomes 2, 3, 4, 7 and 20 that were previously associated with epitasis QTL have been confirmed in a recent fine mapping study using the Virginia chicken model (Pettersson *et al*., 2011). Identifying the genetic mechanism underlying these interactions could considerably advance our understanding of biological evolutionary changes during the selection.

2.2 Next Generation Sequencing

The power of next generation sequencing technology has been widely used to address the diverse biological issues with unprecedented progress. The impressive throughput of the new sequencing technology is providing rapid improvement in areas of the analysis of genome alterations, gene expression and DNA modifications, as well as a constantly growing number of other genetic applications.

NGS technologies include a number of methods that are distinguished basically according to template sequence preparation, the length of reads, and data analysis. NGS can provide higher throughput, accuracy, and a reduced cost compared with previously prevalent Sanger library sequencing. Shotgun, paired-end and mate-pair procedures are often used to further increase the NGS efficiency.

2.2.1 Shotgun Library Sequencing

The Shotgun library sequencing was first developed from the classical Sanger sequencing method, which was the most advanced technique for genomes sequencing between 1995 and 2005. The strategy in shotgun is still applied in the current next generation sequencing such as 454, Illumina and SOLID sequencing, but produces shorter reads (25 – 500 bp) with many hundreds of thousands or millions of reads in a relatively short time. (Karl *et al*., 2009) Thus, the coverage of target sequences increased dramatically in comparison with Sanger

sequencing, but the assembly process is much more computationally expensive.

In Shotgun sequencing library preparation, the amplified genomes are first sheared into many small pieces (25 - 500bp) at random. The cutting fragments contained different ends. With enough coverage, the overlapping ends of different fragments are able to assemble into a continuous sequence across the entire the genome (Figure 2).

The new Shotgun sequencing of NGS, uses the parallel sequencing system with significantly more coverage and greater throughput than traditional Sanger sequencing, enabling sequencing of a whole genome in a relatively short time with unprecedented accuracy (Metzker and Michael, 2010).

The chicken genome resequencing data used in this thesis were first generated by shotgun sequencing, where 35 bases per read with ~5X depth coverage in AB SOLID one fragment library was performed for each chicken line.

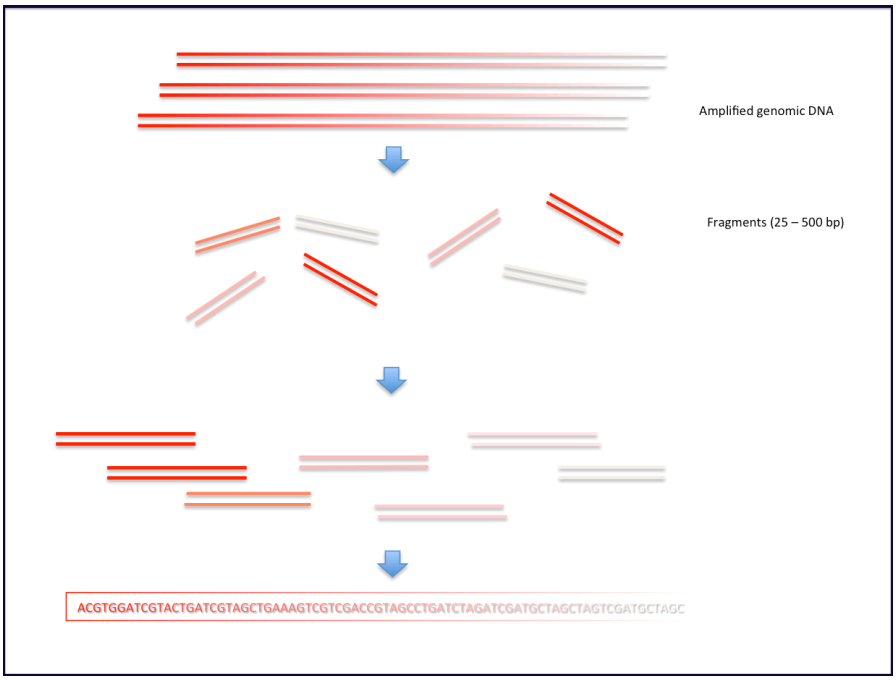

Figure 2. Schematic overview of Shotgun sequencing (based on the figure at "http://en.wikipedia.org/wiki/Shotgun_sequencing").

2.2.2 Paired-end library sequencing

Paired-end reads are a pair of fragments coming from each end of a stretch of DNA sequence. The distance between the two ends is defined by the user allowing flexible sizes, for example, Illumina Genome Analyzer sequencing (200 - 500 bp). In addition to sequence information, the high precision alignment is provided by both reads with the long-range positional information.

One of the prevalent platforms to generate paired-end sequencing library is Illumina Genome Analyzer. In the Illumina paired-end sequencing library preparation, the amplified genome is first sheared into multiple fragments in 200-500 bp at random. Next, the fragments with attached adaptors are denatured to single strands DNA and hybridized to the flow cell panel, where the captured sample DNA is undergone the bridge amplification to generate a clonal cluster.

Next, PCR primers (SP1 and SP2) were attached to each DNA fragments and provide sites for known sequence for sequencing process by synthesis (Figure 3). The sequencing is preformed by four kinds of nucleotides (ddATP, ddGTP, ddCTP, ddTTP), which contain different cleavable fluorescent dye. The final libraries consist of two short DNA segments with originally designed distance (200-500 bp).

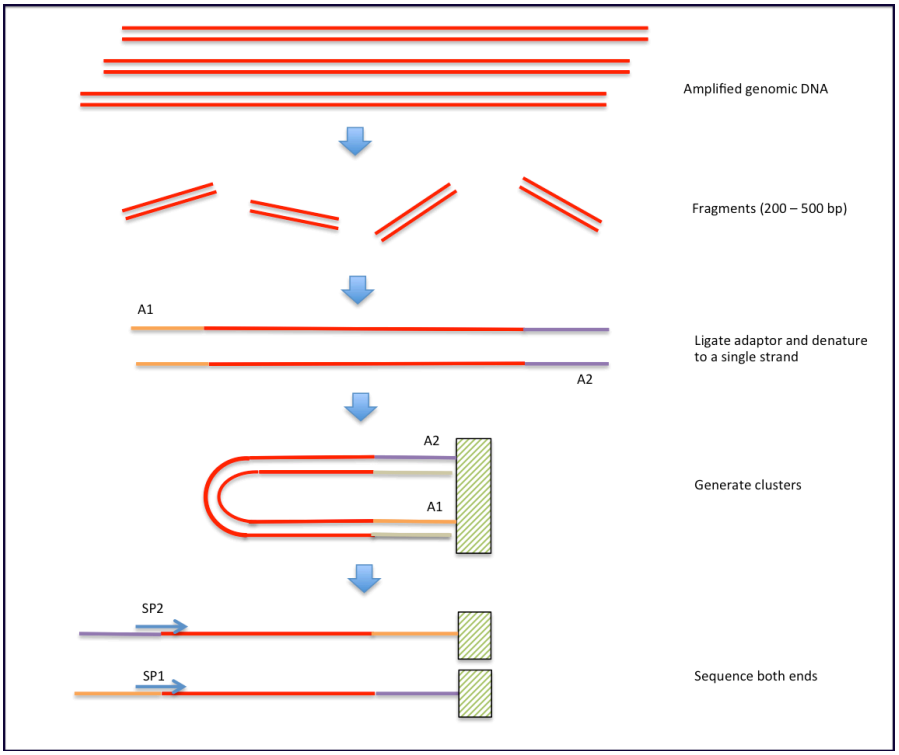

Figure 3. Schematic overview of Paired-End sequencing in Illumina (based on the figure at http://professionalfarmer.github.io/blog/page2/)

2.2.3 Mate-Pair library sequencing

Mate-pair library sequencing is also called long-insert pair-end sequencing, which generates the libraries with insert from 2 kb to 5 kb in size. It is useful for a number of applications such as De Novo sequencing and structural variant detection and can also be more informative than the standard paired-end sequencing due to its long insert size.

One of the typical platforms to generate mate-paired library is Illumina Genome Analyzer. In the Illumina mate-paried library preparation, the amplified genome was sheared randomly into fragments in size 2-5 Kb (Figure 4). Next, the fragments are labelled with biotin at each end and then circularized. The non-circularized fragments will be cleared away by digestion. The circularized fragments are sheared into segments with size-selection (400-600 bp). The DNA segments ligated with biotin labels, which are corresponding

to the ends of the original DNA fragments, are affinity purified. The labelled DNA segment is then added to the adapter sequences and denatured to single strands DNA that subsequently were used to perform the steps of generating clonal cluster and resequencing with tailed PCR primers (SP1 and SP2) (described above in paired-end section 2.2.2). The final libraries contain two short DNA segments with originally separated by several kb.

In this thesis, the part of genome re-sequencing was carried out by using mate-pair libraries, where 50 bases per read with ~7X depth coverage was performed in each chicken line.

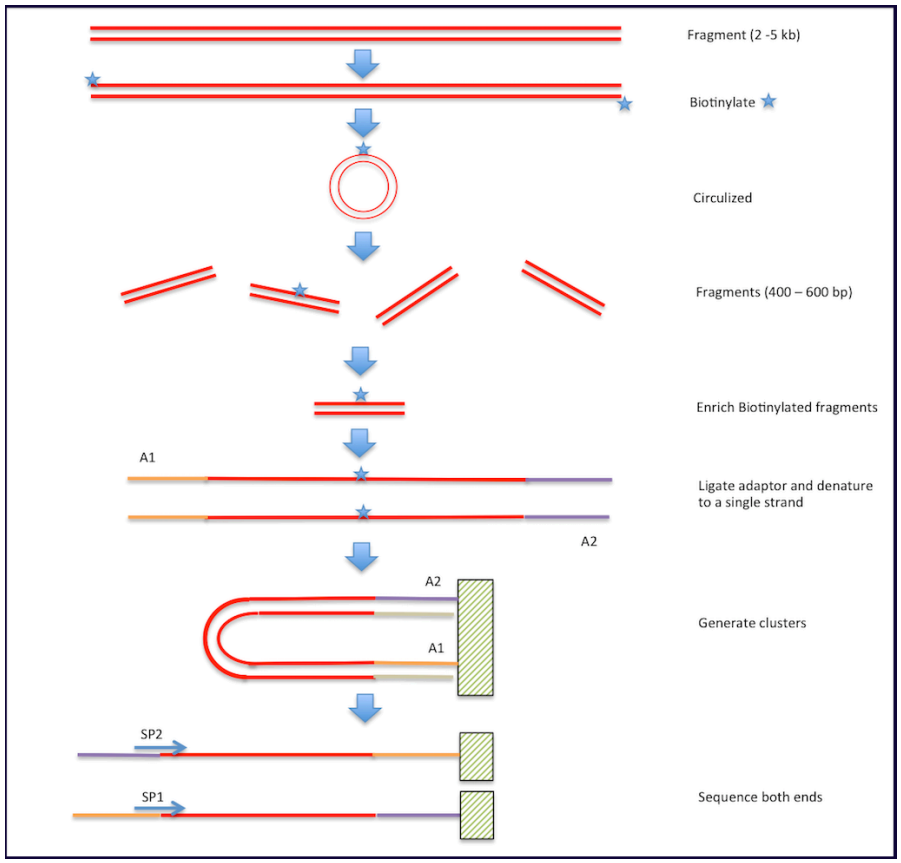

Figure 4. Schematic overview of Paired-End sequencing in Illumina (based on the figure at http://professionalfarmer.github.io/blog/page2/).

2.3 Protein Analysis

Proteins are large biological molecules as a major source of energy to participate in every biological processes including gene regulation, signal transduction, replicating DNA, catalyzing metabolic reaction and etc.

2.3.1 Predicting the Effects of Amino Acids Substitutions

The primary distinction between proteins is based on their sequence of amino acids, which usually initiate the folding of a protein into a specific three-dimensional structure that is critical for the protein function. Recently, the development of new tools for whole genome sequencing, such as NGS technologies, has resulted in a rapid progression of the identification of nsSNPs that cause amino acid substitutions (AAS). AAS have been proven to have a large impact on the corresponding protein and be associated with most of genetic variances known to cause disease in human and other species (Hamosh *et al.*, 2005). However, experimentally characterizing the impact of AAS on protein function is time-consuming and expensive. Thus, a computational method would be a feasible option to help research prioritize AASs for additional study.

Many of the available AAS predictions methods are based on the assumption that mutations influencing the corresponding protein function are highly likely to occur at evolutionarily conserved sites. Widely applied tools are, for example, SIFT (Ng and Henikoff, 2002; Ng and Henikoff, 2003; Ng and Henikoff, 2003) and PolyPhen (Stitziel *et al.*, 2003; Stitziel *et al.*, 2004). SIFT is based on sequence conservation and position-specific scoring matrices with Dirichlet priors, whereas PolyPhen uses sequence conservation and protein ternary structure to model amino acid substitution sites combined with SWISS-PROT annotation (Ng and Henikoff, 2006). However, for evolutionary-distant proteins, the degree of sequence conservation is difficult to evaluate due to a limited amount of orthologous sequences, whereas alternative sources such as physicochemical properties of amino acid could be valuable for functional prediction of AAS.

In this thesis, we developed PASE, a software for predicting the effect of AAS on the hosting protein based on the changes of physico-chemical properties. For this purpose, seven physico-chemical properties were imported from the AAindex database based on a previous study (Rudnicki *et al.*, 2004), which gives every amino acid a unique profile. AAindex is the numerical indices database, where

various physico-chemical and biochemical properties of amino acids are collected from experimental characterization and published literature and represented by a set of numerical values (Kawashima *et al*., 2008). Moreover, the information obtained from PASE can also be combined with knowledge about sequence conservation to further improved functional predictions.

2.3.2 Prediction of Protein functions

Rapid development of high throughput sequencing technologies, such as NGS, has resulted a large number of newly identified proteins. Many protein sequences have never been experimentally characterized their cellular function due to the laborious and expensive wet-lab experiment. However, further development of computational methods for protein sequence analyses can enable protein predictions with increased efficiency.

Protein function prediction methods are consistently based on the information of data-intensive sources, including amino acid sequence homology, protein domain structures, text mining of publications, phylogenetic profiles, and protein-protein interactions. Thus, methods of protein function prediction could be sorted into the following four categories: (1) homology-based function prediction, (2) structure-based function prediction, (3) genomic context-based function prediction and (4) sequence motifs-based function prediction.

The most widely applied method is homology-based function prediction, such as BLAST, which has conventionally relied on detecting similarities between a query sequence (new sequence) and protein sequences in databases with known functions and uses statistical score as an assessment of the validity of the prediction. However, this approach has a failure rate between 20 - 40 % in de novo sequenced genomes (Letovsky and Kasif, 2003).

The structure-based function prediction is based on 3D protein structure similarity between a query sequence and sequences with know function. The Protein Data Bank (PDB) database (Berman *et al*., 2000) is currently the largest protein 3D structure database. The prediction of many programs, such as FATCAT and DeepAlign, are based on scaning an unknown protein structure against the PDB database and report similar structure (Ye *et al*., 2004). However, due to the inadequate sources of protein 3D structure, the application of this approach is inherently limited.

The genomic context-based function prediction relies on some types of correlations between a query protein and proteins with annotated functions. Many of these methods are based on the observation of proteins sharing the same pattern. For example, two or more proteins sharing the same signal transduction pathway in many different genomes are most likely to have a functional link (Sleator *et al*., 2010; Eisenberg *et al*., 2000).

The sequence motifs-based method is based on finding annotated protein motifs within a query sequence, providing new evidence for similar functions. A protein motif, shorter signatures known as motifs, is conserved and repeating patterns in amino acid sequences that are presumed to have an independent biological function.. The development of protein motifs databases such as Pfam (Protein Families Database) (Finn *et al*., 2010) and PROSITE (database of protein domains, families and functional site) (Sigrist *et al*., 2010) has facilitated the application of this approach. The development of the Profat software for identification of functionally related proteins based on repeat motif sequences is presented as a part of this thesis. Once such a group of functionally related proteins are identified, the evolutionary distance of the non-motif part of their sequences is computed and the function of the unknown protein assessed. Profat can be applied to the proteome-wide studies, which can be more sufficient than tools requires 3D protein structures. Profat also has a broad applicability to evolutionary-distant species, where homology-based function prediction may not be effective due to the limited homology sources.

2.4 Parallelizing computing

Parallelizing computing is a form of high performance computing, where many calculations are performed simultaneously. The principle of parallelizing computing is that large tasks can often be divided into smaller segments, which are then done concurrently - in parallel. The parallelizing computing has a wide application in bioinformatics areas, including analyzing the data of NGS, BLAST searching and massive biology data mining.

2.4.1 Multiple-Thread Computing

Multiple-thread computing is one of rudimentary forms of parallelizing computing. Multithreading is the ability of a program or an operating

system process to handle multiple requests without having multiple copies of the running program in the computer.

On a single processor computer, the multi-threading is generally implemented by time division, where the processor switches between threads at any given time. AS this switching action occurs extremely frequently, the user perceives the threads or tasks as running at the same time. On a multiple processor computer, threads processing can be truly synchronal, where every processor or core executes a separate thread simultaneously. Computer with tens of thousands of "off-the-shelf" processors (Figure 5) are conventionally called the supercomputer. It has been widely applied to perform massively parallel computing across many application domains in science including quantum mechanics, applied physics, molecular modeling, mathematical modeling and bioinformatics.

In this thesis, we used two supercomputers (Lindgren in Royal Institute of Technology and UPPMAX in Uppsala University) to perform the parallelizing computing to evaluate the prediction accuracy of Profat (Paper IV) and part of analyzing re-sequencing data of NGS (Paper I and II).

Figure 5. Swedish fastest computer - Lindgren. No.9 in Europe and Nr. 31 worldwide on June 2011 Top 500 list (Photo: Xidan Li).

2.4.2 GPU Computing

The computing using a Graphics Processing Unit (GPU) through a Central Processing Unit (CPU) is called GPU computing. GPUs were first introduced by NVIDIA in 2007, which currently is widely used to accelerate performances in many computation-intensive areas , for example science, engineering, and enterprise applications., In comparison with a CPU, a GPU consists of thousands of smaller, more efficient cores, which was designed to managing multiple tasks in parallel (Figure 6).

Several bioinformatics software are already implemented with GPU computing, such as CUSHAW (parallelize aligning the multiple sequences with GPU), GPU-BLAST (BLAST search with GPU) and GPU-HMMER (Parallelized local and global search with profile Hidden Markov models with GPU). Thus, developing the new versions of software listed in this thesis by GPU computing will be of interest in future study, which will maximize the efficiency of performances.

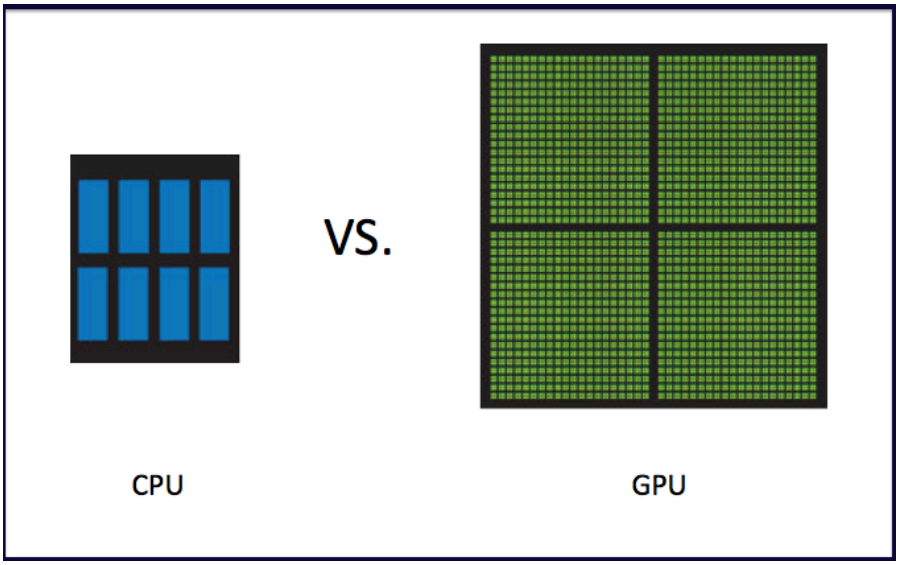

Figure 6. The comparison of a CPU panel with a GPU panel (based on the figure at: http://www.video-stitch.com/gpu-cuda-technical-insights/).

.

3 Aims of the Thesis

The overall aims of the thesis were to:

- use available SNP chip and genome re-sequencing data in previously fine mapped chromosome regions to reveal genetic architecture that contributes to the phenotypic differences between HWS and LWS chicken lines.

- develop efficient computational methods, algorithms and tools for identification of mutations, genes, proteins and pathways underlying phenotypes of interest.

4 Summary of Studies

This thesis is based on studies described in five papers. Paper I and II focus on identifying the genetic variances affecting the phenotypic differences between two divergently selected chicken line on body weight. In the studies presented in the papers III IV and V, my colleagues and I developed three softwares (PASE, Profat and DIPT), and PASE used for the functional prediction of AAS in paper I and II. Profat and DIPT could also be used for functional evaluation of genetics variants. The software developed in thesis was designed to be widely adaptive and could be feasible for many other research projects.

4.1 Paper I

The paper I is the first of the two papers that detects the genetic variances corresponding to the phenotypic differences between HWS and LWS chicken lines. Previous study using an advanced inter-cross line has identified the epistasis pattern between QTL regions on chromosome 2, 3, 4, 7 and 20 (Pettersson *et al*., 2011). The aim of this study is to identify the genetic mechanism underlying these interactions, which would make a valuable contribution to our understanding of evolutionary changes during selection.

4.1.1 Materials and Methods

Mapped epistasic QTL regions to be explored

The epistasis QTL regions were extracted according to the result of previous study (Pettersson *et al*., 2011), where five growth-related QTL regions (Growth2, Growth4, Growth6, Growth9, and Growth12)

were scanned by the same dataset, and each of these regions was performed as a conditional locus in turn. Table 1 shows that the regions demonstrate the epistasis pattern with statistical significance threshold 95%, where the entire population (HWS and LWS) was as the genetic backgrounds.

Table 1. *The coordinates and sizes of the mapped epistasis QTL regions.*

Chromosomes	**Start (Mbp)**	**Stop (Mbp)**	**Size (Mbp)**
2	57.8	60.1	2.3
3	24.5	28.8	4.3
3	28.8	33.9	5.1
3	34.3	39.0	4.7
3	44.5	47.1	2.6
3	49.7	63.1	13.4
3	66.7	68.0	1.3
4	1.4	4.1	2.7
4	6.8	11.9	5.1
7	16.9	37.4	20.5
20	7.1	9.3	2.2

Genome resequencing of pooled population-samples and SNP-calling

Genome resequencing experiments were performed in two separate runs using DNA pools samples from the HWS and LWS lines. The data from two runs were combined to accumulate the re-sequencing coverage in order to promote the accuracy of the SNP identification.

At first run, DNA samples are from two pools of genomic DNA, where each pool included seven males and four females. Genome was re-sequenced using the ABI SOLID one fragment library with 35 bases per read on each pool, where ~5x depth coverage was obtained. At second run, each pool of genomic DNA samples contains eight individuals. The ABI SOLID mate-pair libraries with 50 bases per read were applied, where ~7X depth coverage was achieved in each line. At final, the re-sequencing datasets from the two rounds of sequencing were combined for later SNP calling based on a total of about 12X depth coverage in each line. The MOSAIK software (Lee *et al*., 2013) was applied to perform the sequence assembling, where the RJF genome was used as the reference sequence. The SNP-calling was

implemented by the GigaBayes software, a newer version of PolyBayes (Marth *et al*., 1999).

Identification of candidate SNPs, genes and epistasis pattern in previously mapped QTL regions

At first stage, SNPs between HWS and LWS lines in epistasis QTL region were detected from re-sequencing data. The Variant Effect Predictor (VEP) (McLaren *et al*., 2010) from Ensembl database (Flicek *et al*., 2014) was used to annotate newly discovered SNPs on known Ensembl genes.

Next, identified SNPs on genes were here filtered out with the greatest allele frequency differences. Moreover, for SNPs detected in coding regions as non-synonymous mutations, we used PASE to evaluate their effect to the hosting protein function according to the changes of physico-chemical properties of amino acid (Li *et al*.,2013).

Further, each gene was scanned to see if any occurred in a growth related pathway, using the KEGG pathway database (Kanehisa *et al*., 2014). If two or more genes from different QTL regions occurred in the same growth-related pathway, they were listed as possible candidate genes for the non-additive interaction observed from the QTL study.

4.1.2 Results and Discussion

In total, 230 SNPs in CpG island, 16 SNPs in UTR region, 8 SNPs in splicing sites and 16 Non-synonymous SNPs have been detected, which are located within 11 epistasis patterns.

The candidate SNPs in the QTL region of chromsome 7 have been detected in all 11 epistasis patterns. This result is consistent with previous predictions (Pettersson *et al*., 2011; Carlborg *et al*., 2006), where the QTL in chromosome 7 has interactions with the QTL in chromosome 2,3,4 and 20 and. Moreover, the candidate SNPs from the QTLs in chromosome 3 and 7 (table 3) are involved in 10 of 11 epistasis patterns, which imply that the interaction between chromosome 3 and 7 may contribute to a large effect to the phenotypic differences between HWS and LWS lines.

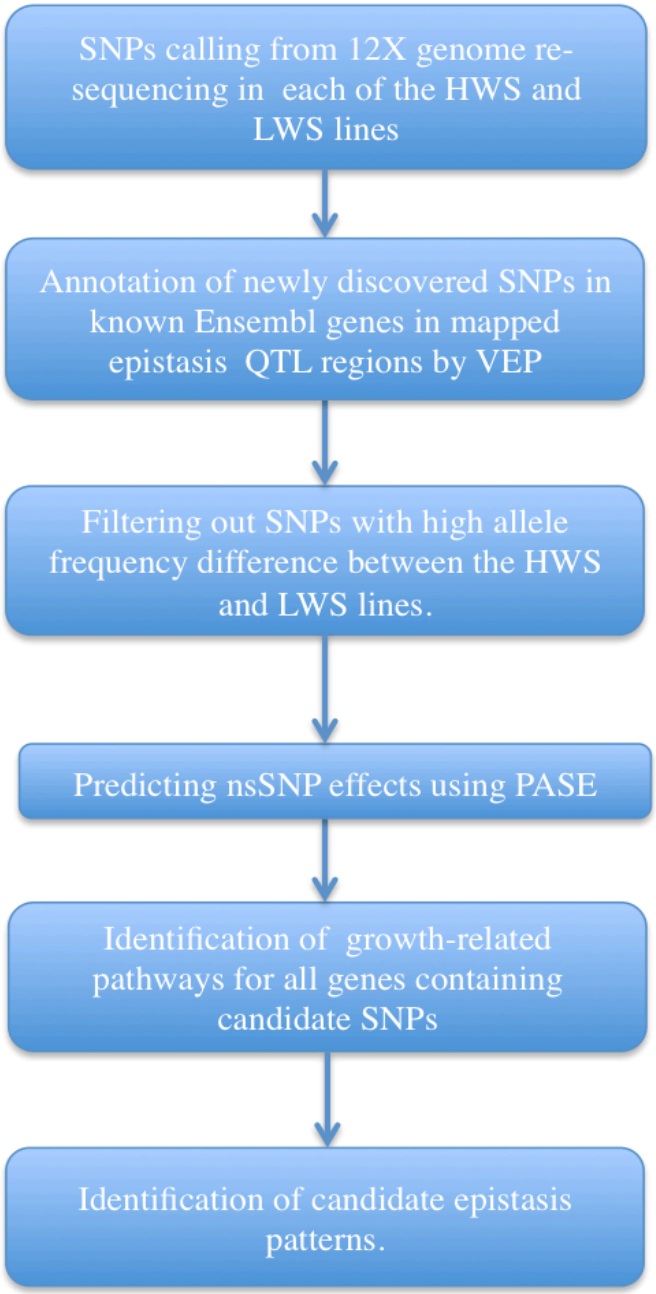

Figure 6. The process to identify the epistasis pattern for phenotypic differences between the HWS and LWS lines.

4.2 Paper II

In addition to Paper I, the study presented in Paper II also revealed genetic polymorphisms which may contribute to the differences of phenotypic traits between two chicken lines that have been divergently selected for body weight at 56 days age.

Previous studies have mapped several QTL regions using an F2 inter-cross between the HWS and LWS lines (Jacobsson *et al*., 2005) and an advanced inter-cross line to confirm and narrow down the QTL regions (Besnier *et al*., 2011). The aim of the study presented in Paper II was to identify genetic elements underlying these QTL regions. In this study, we also developed a bioinformatics strategy to explore already identified QTL regions to identify candidate genes for growth trait in chicken.

4.2.1 Materials and Methods

Mapped QTL regions to be explored

Seven QTLs on chromosomes 1–5, 7, and 20 have been previously mapped on the selection of body-weight trait from generation 41 (Jacobsson *et al*., 2005; Besnier *et al*., 2011) (Table 2*).*

Table 2. Fine-mapped growth QTL regions with significance (Ahsan et al. 2013 and Besnier et al. 2011)

Chromo-some	QTL	Region name	Start (Mbp)	End (Mbp)	Size (Mbp)
1	Growth1	C1G1	169.9	181	11.4
2	Growht2	C2G2	47.9	65.4	17.5
3	Growth4	C3G4	24	68	43.9
4	Grwoth6	C4G6	1.3	13.5	12.1
5	Grwoth8	C5G8	33.6	39	5.3
7	Grwoth9	C7G9	10.9	35.4	24.5
20	Grwoth12	C20G12	7.1	13.8	6.7

Individual genome-wide 60K SNP chip genotyping

Twenty individuals from each of the high and low lines at generation 41 have been genotyped with the 60K SNP chip in previous study (Marklund and Carlborg, 2010). These genotype data were used to estimate the SNPs allele-frequency differences between the HWS and LWS chicken lines.

The genome re-sequencing of pooled population-samples and SNP-calling has been described in Paper I section (see above).

Genetic divergence analysis using the flanking-SNP-value method in re-sequencing data

We used previously developed method the flanking-SNP-value (FSV) (Marklund and Carlborg, 2010) to estimate the genetic divergence between HWS and LWS lines. The principle of FSV is computing the allele frequency differences between two lines for each SNP and its flanking SNPs in both directions within an interval, where SNPs were assumed to have a high degree of linkage disequilibrium. FSV value is computed as,

$$\text{FSV} = \frac{\left(\sum_{i=1}^{N_H}\left|S_{c_i}^{H} - S_{c_i}^{L}\right|\right)}{N_H} \text{ x } \frac{\left(\sum_{j=1}^{N_L}\left|S_{d_j}^{L} - S_{d_j}^{H}\right|\right)}{N_L}$$

where $S_{C_i}^{H}$ and $S_{C_i}^{L}$ indicate the proportion of reads in HWS and LWS lines respectively for SNP at position *c*; likewise, $S_{d_j}^{L}$ and $S^{H}{}_{d_j}$ represent the same component for SNP at position *d;* N_H and N_L are the total number of reads proportion of SNPs within flanking regions in the HWS and LWS lines, respectively.

A combined scores for prioritizing the previously mapped QTL regions

To narrow down the mapped QTL region in order to facilitate identifying the candidate genes and SNPs, we developed a bioinformatics strategy, which was represent as a combined data score (CDS),

$$CDS = \{[(\, FSVscore + SNPchip\ allele\ freq.\,)/2] + (\, Normalized\ score\ of\ QTL\_Model\_B)\}/2$$

where the $FSVscore$ represents the genetics divergence between two chicken lines; $SNPchip\ allele\ freq.$ indicates the allele frequency differences based on the individual 60K SNP chip genotyping; $Normalized\ score\ of\ QTL\_Model\_B$ represents the effect of each SNP in the previously mapped QTL regions.

Identification of candidate genes and SNPs in previously mapped QTL regions

We first used the CDS scores to narrow down previously mapped QTL regions. Next SNPs identified between HWS and LWS lines from 12x

genome re-sequencing were annotated on known Ensembl genes in prioritized QTL regions by VEP (McLaren *et al.*, 2010).

Genes containing candidate SNPs were filtered out with growth-related annotation from DAVID database (Huang *et al.*, 2009). Moreover, we used PASE to evaluate the effect of candidate nsSNPs to the hosting protein function according to the changes of physico-chemical properties of amino acid (Li *et al.*, 2013).

In consequence, several most promising candidate genes and SNPs are identified in the prioritized QTL regions (Table 3).

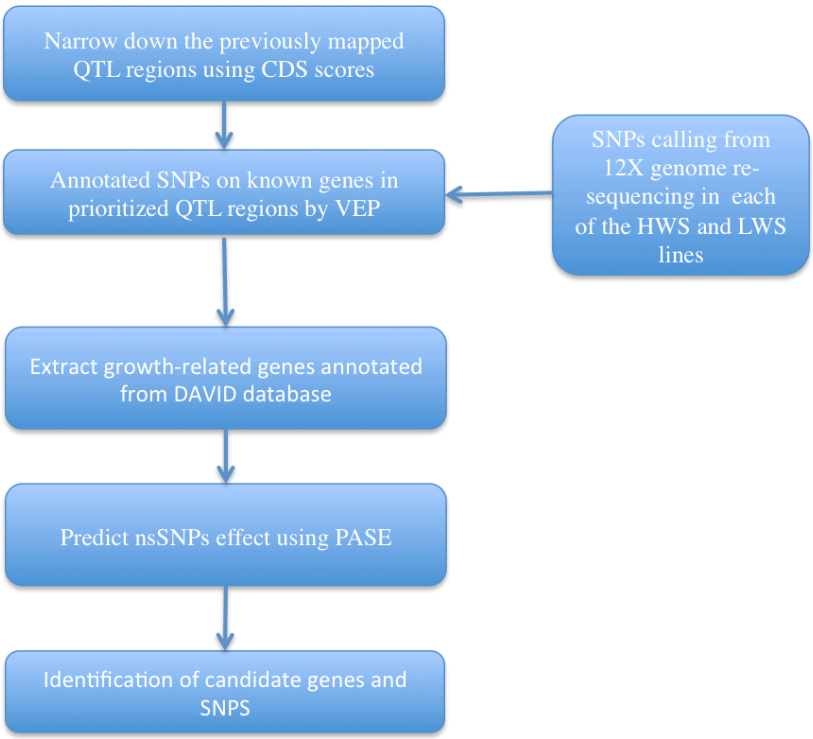

Figure 7. The workflow to identify the candidate genes and SNPs corresponding to the phenotypic differences between the HWS and LWS lines.

4.2.2 Results and Discussion

In this study, we collected several top popular databases, re-sequencing data of NGS and SNP chip, previously developed methods and mapped epistasis QTL regions to systematically evaluate genetic variants based on their presumed effect on gene function and

relevance for growth. In consequence, 10 candidate SNPs are detected to be related to the phenotypic differences between HWS and LWS (Table 3). Three of ten candidate SNPs are detected in QTL of chromosome 7, which implies QTLs in chromosome 7 is the key region to regulate the phenotypic differences. It is consistent with previous findings (Carlborg *et al*., 2006). Moreover, the SNP (22711910 in chromosome 7) in the glucagon was one of strongest candidate mutations. Glucagon has been well described to have effect on appetite (Jacobsson *et al*., 2005), a trait for which HWS and LWS lines show a striking difference.

In conclusion, our results can be used for further verification and experimental evaluation to improve our understanding of genetic regulation of growth-related traits.

Table 3. *Candidate mutations identified in the evaluated QTL regions. (Ahsan et al., 2013)*

Chromosome	SNP (bp)[1]	Gene	SNP location[2]	PC Score[3]	EC Score[4]	PE Score[5]
1	174634021	Asparagine-linked glycosylation 11 homolog (ALG11)	CpG island, upstream	N/A	N/A	N/A
2	63823523	Endothelin 1(EDN1)	CpG island, upstream	N/A	N/A	N/A
3	33678270	Cysteine rich transmembrane BMP regulator 1 (CRIM1)	Protein code, NS K/I	0.67	0.63	0.42
4	12044024	Similar to receptor upstream tyrosine kinase (VEGFR-2)	CpG island,	N/A	N/A	N/A
4	12902414	Fibroblast growth factor 16 (FGF16)	CpG island, downstream	N/A	N/A	N/A
5	38316301	Sorting nexin 6 (SNX6)	CpG island, upstream	N/A	N/A	N/A
7	21686625	Growth factor receptor-bound protein 14 (GRB14)	CpG island, downstream	N/A	N/A	N/A
7	22711910	Glucagon (GCG)	CpG island, downstream	N/A	N/A	N/A
7	24802616	Insulin-like growth factor binding protein 2 (IGFBP2)	Protein code, synonymous, CpG island	N/A	N/A	N/A
20	8715398	Baculoviral IAP repeat-containing 7 (BIRC7)	Protein code, NS I/V	0.29	0.14	0.04

[1]Coordinates based on the Chicken (Gallus gallus) assembly v 2.1/galGal3; [2]Location of the SNP in gene and also amino acid substitution in case of non-synonymous (NS) SNP; [3]Physico-chemical score of AAs calculated by PASE (Li *et al*., 2013); [4]Evolutionary conservation score of AAS calculated using PASE (Li *et al*., 2013); [5]Combined score of PC and EC of AAS calculated using PASE (Li *et al*., 2013).

4.3 Paper III

Paper III presents a development of the novel software PASE (Prediction of Amino acid Substitution Effects) that efficiently can predict the effect of AAS on the hosting protein function based on changes of physico-chemical properties.

With an increased use of NGS technologies such a large number of nsSNPs have been identified. These nsSNPs cause amino acid substitutions, which in many cases are associated with diseases (Hamosh *et al*., 2005). The aim of this study was to develop a software that efficiently predicts amino acid substitution effects on protein structure and function.

4.3.1 Implementation

Each of 20 naturally occurring amino acids has a characteristic profile of physicochemical properties. Therefore, any amino acid substitution may influence the final protein structure by altering its physicochemical properties. Here, to calculate the changes of physico-chemical properties led by AAS, we selected seven physico-chemical properties from the AAindex database (Kawashima *et al*., 2008), which make every amino acid unique profile (Rudnicki *et al*., 2004). An euclidean norm formula was applied to compute the physicochemical property changes of AAS between the original and the substituting amino acid:

$$P = \sqrt[2]{\sum_{7}^{i=1}(aa_o^i - aa_s^i)^2}$$

where aa_o^i and aa_s^i indicate one of the selected seven physicochemical properties of original amino acid and substituted amino acid, respectively.

Moreover, sequence conservation approach has been widely applied to predict the effect of AAS. Highly conserved degree conventionally performs as an indicator to a functional importance of the amino acid, where substitutions tend to be deleterious, whereas those within areas of low conservation are often tolerated. In this study, we also developed a new algorithm to calculate a sequence conservation score based on the multiple sequence alignment (MSA), where the NCBI-blastp is first used to search for functionally related protein sequences and Clustalw is subsequently applied to generate an alignment with multiple homologous sequences. Thus, the MSA conservation (MSAC) score at each site is calculated by using the following formula,

$$C = r\,(1 - 0.95^N)$$

where N is the number of sequences of MSA, and r is the proportion of the amino acids of interest in the same column of MSA(Figure 8). The formula $(1 - 0.95^N)$ (Pei *et al.*, 2001) indicates the probability of 20 different amino acids in a position for N random equal frequent amino acid sequences. For example, when N = 1, the probability of each amino acid in a position is 0.05, which is consistent with 20 amino acids with frequency 1/20 each.

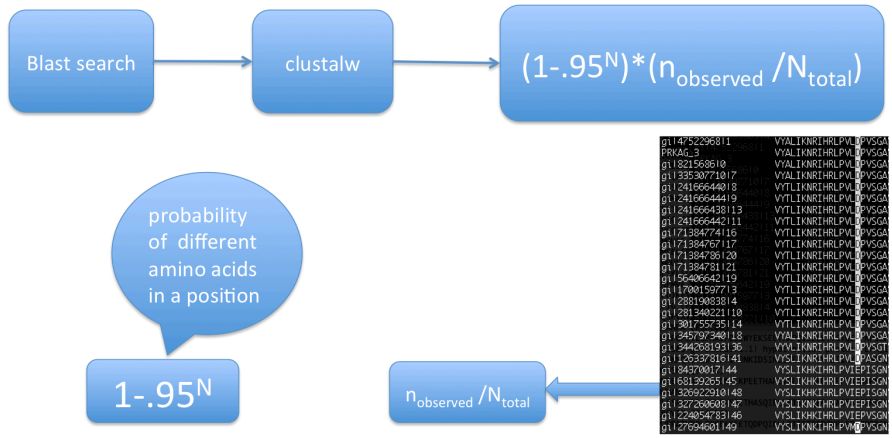

Figure 8. Schematic overview of calculation of sequences conservation score

For the advanced prediction, the PASE score can be combined with sequence conservation score (MSAC) by creating a combined score – PASEC, which is computed as following formula,

$$S = PC$$

where S indicates PASEC score, P is the score of physico-chemical properties changes and C is sequence conservation score (MSAC). The score ranges from 0 to 1, where 0 is neutral, and higher ratio indicate stronger predicted effects on the protein.

4.3.2 Results and Discussion

To verify the accuracy of PASE prediction, we tested with 3338 AASs with functional effects previously predicted with the widely used tools SIFT and POLYPHEN in Ensembl database. A total of 1978 and 1637 AASs were predicted as “tolerated” and “benign” with average score 0.39 and 0.37 by SIFT and PolyPhen, respectively, whereas 1351 AASs predicted as “deleterious” by SIFT and 1162 AASs predicted as

"probably damaging" by PolyPhen showed the average score 0.51 and 0.53 (Table 4). The Distributions of different scores were shown in Figure 9, where PASE MSAC and PASEC scores were shown a consistent trend with the classifications of SIFT and POLYPHEN.

Table 4. *Predictions of nsSNPs from Human Chromosome 22.*

Name of tools	Classifications	Number of AAS	MSAC Conservation score	PASE Physicochemical score	PASEC scores
SIFT	tolerated	1987	0.47	0.39	0.18
	deleterious	1351	0.6	0.51	0.3
PolyPhen	benign	1637	0.44	0.37	0.16
	possibly-damaging	539	0.56	0.43	0.24
	probably-damaging	1162	0.63	0.53	0.33

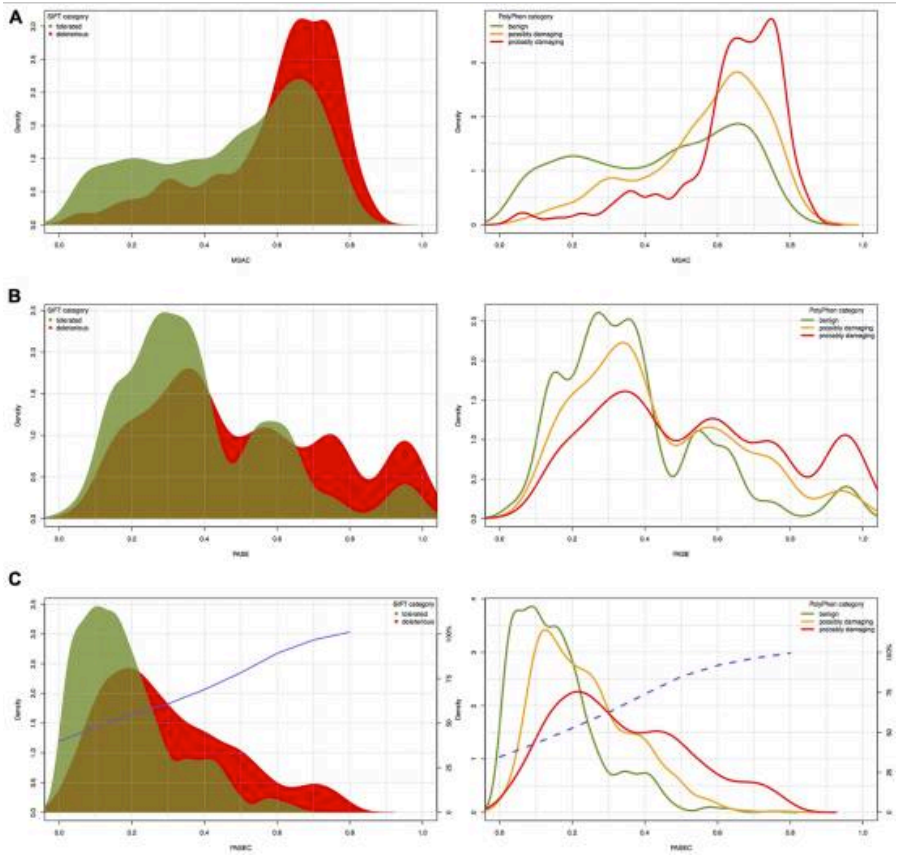

Figure 9. The distribution of (A) PASE, (B) MSAC, and (C) PASEC scores within different SIFT and PolyPhen prediction classes. Blue solid and dashed lines in panel (C) correspond to the probability of deleterious/damaging prediction from AAS's PASEC scores.

4.4 Paper IV

Paper IV is another software development paper, where we present the novel software Profat (Protein function annotation tools) that predicts the function of unknown proteins by using a motif-profile based analysis, where repeat motifs sequences are first used for protein family classification and later based on a subsequent phylogenetic analysis on non-motif sequence similarities within protein families.

An increased use of NGS technologies has provided a large number of identified novel protein sequences that needs to be annotated. Experimental characterizations of the function of these proteins are

often time-consuming and expensive. The aim of this study was to develop bioinformatics tools that can facilitate annotation of novel proteins.

4.4.1 Implementation

Profat applies the protein motif profile as an indicator to classify the query sequence into a motif-containing protein family, where the proteins are presumed to have similar functions. Next, the evolutionary distance of the non-motif part of their sequences is computed, where sequence could not be predicted based on available motif knowledge i.e. the so called non-motif part of the protein sequence. To achieve this, the motifs in each protein sequence in the motif-containing protein family, as well as the query sequence, are identified by using the HMMER package (Richard *et al*., 1998) and subsequently filtered out by replacing them with multiple 'X'-placeholders. BLAST (Altschul *et al*., 1998) is then used to query each filtered-out sequence against the database of the same set of filtered-out sequences to identify similar sequences of the non-motif part. The evolutionary distance between two filtered-out sequences is subsequently computed using the following formula:

$$ED = 1-(s/(t1+t2)/2)$$

where t1 and t2 are the total number of hits (i.e. indicators of the sequence similarity) of protein sequence 1 and protein sequence 2, respectively, and s is the number of hits shared by the two proteins sequences. Further neighbor-joining (Saitou *et al*., 1987) clustering is used to build the phylogenetic tree that can visualize the evolutionary distance between the query sequence and other sequences (Figure 10). The prediction is based on finding the evolutionary distance between a query sequence and protein sequences in databases with known functions.

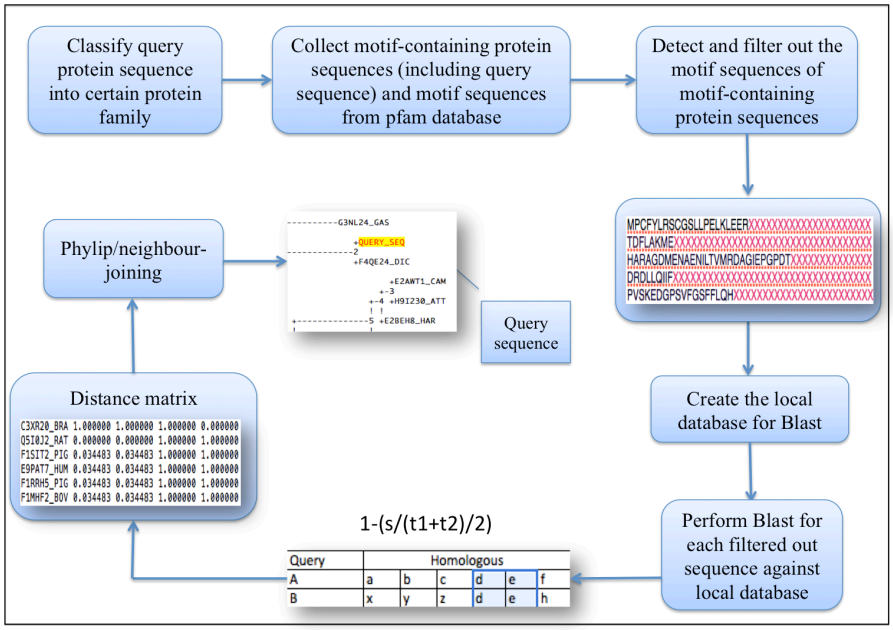

Figure 10. Schematic overview of the Profat algorithm.

4.4.2 Results and Discussion

To explore the predictive ability of Profat, we selected two representative protein sequences as examples.

gi:410910826 – a characterized protein

The protein "gi:410910826" in the Fugu rubripes genome is a previously described as partitioning defective 6 homolog gamma-like protein (Joberty *et al.*, 2000) in NCBI protein annotation database. Here, Profat has predicted it also as a partitioning defective 6 homolog gamma protein involved in the process of cell division, which is consistent with the annotation in NCBI database.

gi:328865585 - an evolutionary-distant protein

Many protein predictions methods, such as BLAST, consistently rely on the sequence similarity, but are often impotent to evolutionary-distant proteins due to the inadequate resources of orthologous sequences. However, Profat includes the motif structure characterization in the prediction process, which is better suited to analyze such proteins.

Protein "gi:328865585" is from Dictyostelium fasciculatum, a species of soil-living amoeba, which is unable to be predicted in BLAST

due to the lack of orthologous sequences. By using Profat, "gi:328865585" first was classified to P31comet-containing protein family, and subsequently predicted to be as MAD2L1-binding protein-like and suggested to be involved in the regulation of exit from mitosis (Yang *et al*., 2007).

Moreover, the structural similarity is more closely related to the protein function than the sequence similarity, and similar sequence does not always imply the similar function. Profat uses protein motif characterization as an indicator to perform prediction. Thus, the prediction of profat is more convincible.

4.5 Paper V

Paper V is another software development paper, where we present a novel software DIPT (Detecting Interactions and Pathways Tool) that identifies candidate causative genes underlying an expression phenotype by searching the QTL/GWA studies regions corresponding to expression phenotype data.

An increasing application of new tools for detecting genetic variances like SNP arrays and NGS technologies have resulted in an accelerated number of GWA studies or QTL mapping analyses (Flint *et al*., 2012; Pettersson *et al*., 2011; Carlborg *et al*., 2006). Additionally, the high throughput phenotyping of RNA expression, using micro-arrays or real-time PCR, have led to the large number of expression phenotype data (Maskos and Southern, 1992). However, the approach to find functional interactions between expression phenotypes and QTLs is scarce. The aim of this study was to develop a bioinformatics tool to address the lack of high throughput analysis tools that combine association studies with expression phenotype data.

4.5.1 Implementation

DIPT creates a link from expression phenotypes to genotypic data with current information in bioinformatics databases. At first stage, DIPT collects the genes from Ensembl (Flicek *et al*., 2013) in the flanking region of the QTL corresponding to the phenotype genes. Meanwhile, interaction database BioGRID (Chatr-Aryamontri *et al*., 2013) was imported to retrieve all the genes known to interact with the expression phenotype gene. Next, genes detected in both queries (i.e. the flanking region of the QTL and the interaction database) are

presented as candidate genes affecting the expression phenotype (Figure 11).

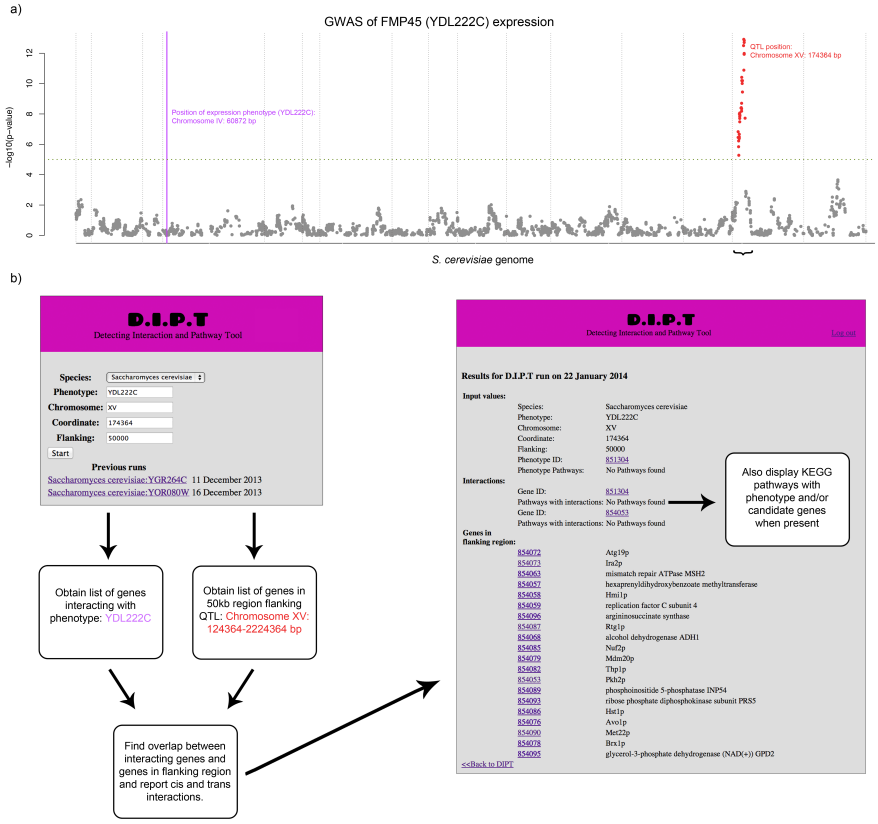

Figure 11. Schematic overview of DIPT. (a), the illustration of the expression phenotype data and the QTL position; (b) the working process of the strategy applied in DIPT.

4.5.2 Results and Discussion

In this study, we tested a publically available yeast dataset to explore the predictive ability of DIPT, where genotypes and expression phenotypes on individuals were developed from crossing between a laboratory S. cereviciae strain (BY4716 , isogenic to S288C) and a wild isolate (RM11-1a) (Brem *et al.*, 2005; Smith *et al.*, 2008; Storey *et al.*, 2005). The dataset contains 109 haploid segregants, each of which was cultivated in two conditions with either glucose or ethanol as the main carbon energy source. The expression profiles for 4482 genes were identified for each segregant in both treatments. For each segregant, a set of 2956 SNP markers were genotyped, which provide an average

marker density of one marker per 4.1 kbp for QTL further analysis. In total, 8387 eQTL were identified for the 4482 expression phenotypes across the two treatments (Nelson *et al.*, 2013).

Each of the 8387 eQTL was analyzed with DIPT to search causative genes to the corresponding expression phenotype. We selected the different flanking region sizes 30000, 50000 and 100000 bps to perform the analysis separately. As a result, an average number of 0.39, 0.59 and 1.13 causative genes per eQTL region were identified respectively (Figure 12).

With screening known interaction approach, DIPT has significantly narrow down the number of candidate genes in eQTL region. However, it is important to realize that the size of flanking region is very sensitive to the result, as some causative genes may be distant from the eQTL position or some genes unrelated to the eQTL region of interest.

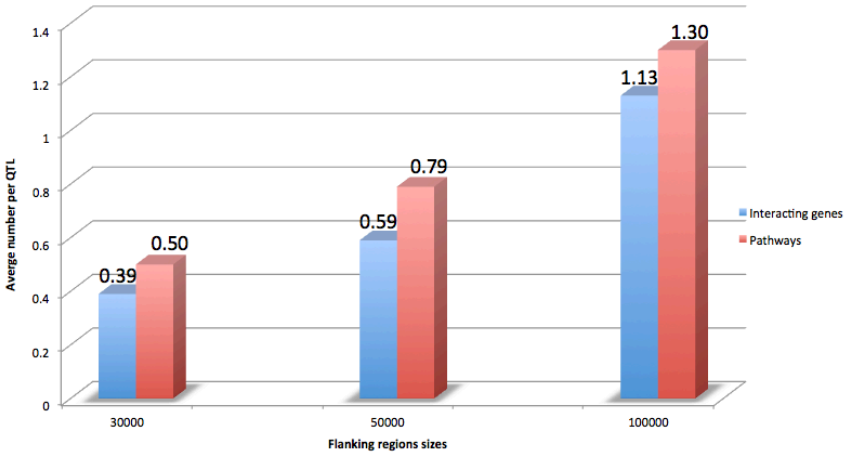

Figure 12. Predict causative genes for 4482 expression phenotypes from Saccharomyces cerevisiae. The blue and red columns indicate the average number of interacting genes and pathways per QTL, respectively.

5 Future research

5.1 Experimental characterization of identified genetic variants

In paper I – II, we searched the genetic variation underlying the QTL effects on chicken body weight. We described a computational approach where different datasets were combined to identify the most promising candidate genes and SNPs. However, we did not confirm our predictions by any other experimental performance. Thus, experimental evaluation of these genetic variants would be the further research plan, which could provide evidence for the causative mutations with effects on growth which could be highly for our understanding of genetic regulation of growth-related traits .

Moreover, our described approach could also be useful in many other projects where causative mutations within chromosomal regions need to be traced and identified.

5.2 High performance computing

A rapid growth of NGS technologies in recent years has lead to an explosive data growth in biomedicine. However, processing and summarizing the large amount of data have been hampered by inefficient computing. The three softwares developed and described in this thesis can facilitate efficient, robust and reproducible data analysis workflows.

Recently, due to the extensive functionality in scientific computation, the application of GPU has been rapidly accelerated. With hundreds of processors, GPU improves the significantly stronger efficiency and performance especially in the data-intensive experiment. Thus, developing the GPU version of softwares listed in this thesis would be near future project, which would maximize the

efficiency of application and provide an all-new set of insights to exploit the genetic data, for example, the re-sequencing data from NGS technologies.

6 Conclusions

6.1 Paper I and II

We combined resequencing data of NGS, SNP chip genotyping and bioinformatics analysis to systematically evaluate genetic variants based on their presumed effect on gene function and relevance for growth at previous mapped QTL regions. A list of epistasis patterns, transcripts and SNPs have bee identified to be related to the phenotypic differences between HWS and LWS, which will facilitate further verification and experimental evaluation. This could help to unravel and understand more about genetic regulation of growth-related traits as well as other complex traits, which may benefit animal breeding, human medicine and/or other areas of biology.

6.2 Paper III

Here, we present a new software, PASE that used the new strategy to predict the effect of AAS to its corresponding protein function. Compared with other method like SIFT and PolyPhen based on the degree of sequence conservation, PASE predicts the effect of AAS by calculating the changes of physicochemical properties of amino acids, which is very effective to the AAS of evolutionary-distant protein. In this study, we also invent a new formula to calculate the degree of sequence conservation. Moreover, we demonstrated that the score based on both sequence conservation and physicochemical properties is a useful way to increase the functional prediction accuracy. PASE complements other tools and facilitates to prioritize the most promising mutations among the large amount of candidate mutations for

a phenotype of interest, which will help to identify the genetic factors underlying complex traits.

6.3 Paper IV

The Profat software is based on a new strategy for functional annotation of novel proteins, which is very effective for evolutionary-distant proteins in comparison with other tools based on sequence similarity. As using motif profile analysis approach, Profat can be applied to the proteome-wide study, which is proposed as an alternative to sequence based strategies to predict function of novel gene products when 3D data is not available.

6.4 Paper V

The DIPT software can help us reveal the genetic mechanisms underlying an expression phenotype by searching genetic variants from QTL/GWA studies regions corresponding to the expression phenotype. Results of DIPT could help prioritizing the most promising genes for further experimental investigation.

Reference

Ahsan, M., Li, X., Lundberg, A.E., Kierczak, M., Siegel, P.B., Carlborg, O. and Marklund, S. (2013). Identification of candidate genes and mutations in QTL regions for chicken growth using bioinformatic analysis of NGS and SNP-chip data. *Front Genet*. 4: 226.

Altschul. S., Gish, W., Miller, W., Myers, E., Lipman, D. (1998). Basic local alignment search tool. *Journal of Molecular Biology*, 215 (3): 403–410.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov,

Besnier, F., Wahlberg, P., Rönnegård, L., Weronica, EK., Andersson L., Siegel P.B., et al. (2011). Fine mapping and replication of QTL in outbred chicken advanced intercross lines. *Genet. Sel. Evol*. 43 3.10.1186/1297-9686-43-3

Boa-Amponsem, K., Dunnington, E.A. & Siegel P.B. (1998). Diet and humoral responsiveness of lines of chickens divergently selected for antibody response to sheep red blood cells. *Avian Dis*, 42, 565-71

Brem, R.B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A*, 102: 1572–1577.

Carlborg, O., Jacobsson, L., Ahgren, P., Siegel, P.B., Andersson, L. (2006). Epistasis and the release of genetic variation during long-term selection. *Nature Genetics* 38:418-20.

Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., et al. (2013)The BioGRID Interaction Database: 2013 update. *Nucleic Acids Res*; 41(D1):D816-D823.

Clutton-Brook, J. (1995). A natural history of domestic animals. *Cambridge: Cambridge Univ Press. Second edition.*

Dunnington, E.A,, and Siegel, P.B. (1985). Long-term selection for eight-week body weight in chickens-direct and correlated responses. *Theor. Appl. Genet.* 71:305.

Dunnington, E.A., Siegel, P.B. (1996). Long-term divergent selection for eight-week body weight in White Plymouth Rock chickens. *Poultry Sci* 75: 1168–1179.

Eisenberg, D., Marcotte, E.M., Xenarios, I., Yeates T.O. (2000). "Protein function in the post-genomic era". *Nature* 405 (6788): 823–826. doi:10.1038/35015694. PMID 10866208.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., et al. (2010). The Pfam protein families database. *Nucleic Acids Res* 38: D211–222. doi:10.1093/nar/gkp985. PMC 2808889. PMID 19920124.

Flicek, P., Ahmed, I., Ridwan, A.M., Barrell, D., Beal, K., Brent, S., et al. (2013). Searle Ensembl 2013. *Nucleic Acids Research,* 41 Database issue:D48-D55 doi:10.1093/nar/gks1236

Flint, J., and Eskin, E., (2012). Genome-wide association studies in mice. *Nature Reviews Genetics*, 13, 807-817.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A. (2005) Online Mendelian Inhertiance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 2005 33:D514–17.

Huang, D.W., Sherman, B.T., Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nuclei Acids Res.* 37 1–13.10.1093/nar/gkn923

I.N., Bourne P.E. (2000). "The Protein Data Bank". Nucleic Acids Res 28 (1): 235–242. doi:10.1093/nar/28.1.235. PMC 102472. PMID 10592235.

Jacobsson, L., Park, H.B., Wahlberg, P., Fredriksson, R., Perez-Enciso, M., Siegel, P.B., et al. (2005). Many QTLs with minor additive effects are associated with a large difference in growth between two selection lines in chickens. *Genet. Res*. 86 115–125.10.1017/S0016672305007767

Joberty, G., Petersen, C., Gao, L., Macara, I.G. (2000). The cell-polarity protein Par6 links Par3 and atypical protein kinase C to Cdc42. *Nat Cell Biol*., 2(8): 531-539.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205.

Karl, V, et al. (2009). "Next Generation Sequencing: From Basic Research to Diagnostics". *Clinical Chemistry* 55 (4): 41–47. doi:10.1373/clinchem.2008.112789

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*. 36, D202-D205 (2008).

Lee, W.P., Stromberg M., Ward A., Stewart C., Garrison E., Marth G.T. (2013). MOSAIK: a hash-based algorithm for accurate next-generation sequencing read mapping. *arXiv preprint arXiv*: 1309.1149.

Letovsky, S, Kasif, S. (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics,* 19: i197-i204

Li, X., Kierczak, M., Shen, X., Ahsan, M., Carlborg, ö., Marklund, S. (2013). PASE: a novel method for functional prediction of amino acid substitutions based on physicochemical properties. *Front. Genet.* 4:21.10.3389/fgene.2013.00021

Marklund, S., Carlborg ö. (2010). SNP detection and prediction of variability between chicken lines using genome resequencing of DNA pools. *BMC Genomics* 11 655.10.1186/1471-2164-11-665

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z.J., Zakeri, H., et al. (1999). A general approach to single-nucleotide polymorphism discovery. *Nat. Genet*. 23 452–456.10.1038/70570

Maskos, U. and Southern, E.M. (1992) Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ.*Nucleic Acids Res* 1992, 20 (7): 1679–84.

McLaren, W., Pritchard, B., Rios, D., Chen, Y.A., Flicek, P., Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. Bioinformatics 26 2069–2070.10.1093/bioinformatics/btq330

Nelson, R.M., Pettersson, M.E., Li, X. and Carlborg Ö. (2013). Variance heterogeneity in Saccharomyces cerevisiae expression data: trans-regulation and epistasis. *PLoS ONE* 8(11): e79507.doi:10.1371/journal.pone.0079507

Newmyer, B.A., Nandar, W., Webster, R.I., Gilbert, E. and Siegel, P.B. (2013) *Behav Brain Res,* 236:327.

Newmyer, B.A., Siegel, P.B. and Cline, M.A. (2010) Neuropeptide AF differentially affects anorexia in lines of chickens selected for high or low body weight. *J. Neuroend.* 22:1-6.

Ng, P.C., Henikoff, S. (2002). Accounting for Human Polymorphisms Predicted to Affect Protein Function. *Genome Res*, 12(3):436-46.

Ng, P.C., Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*; 31(13):3812-4.

Ng, P.C., Henikoff, S. (2006). Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annu Rev Genomics Hum Genet*, 7:61-80.

Noble, D.O., Dunnington, E.A. and Siegel, P.B. (1993). "Ingestive behavior and growth when chicks from lines differing in feed consumption are reared separately or intermingled." *Appl. Anim. Behav. Sci.* 35:359-368.

Parmentier, H. K., Nieuwland, M. G., Rijke, E., De Vries Reilingh, G. & Schrama, J. W. (1996). Divergent antibody responses to vaccines and divergent body weights of chicken lines selected for high and low humoral responsiveness to sheep red blood cells. *Avian Diseases* 40, 634 – 644.

Pei, J., Grishin, N.V. (2001). AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics,* 17(8): 700-12.

Pettersson, M., Besnier, F., Siegel, P.B., Carlborg, Ö. (2011). Replication and explorations of high-order epistasis using a large advanced intercross line pedigree. *PLoS Genet*.7, e1002180.10.1371/journal.pgen.1002180

Pinard van der Laan, M., Siegel, P.B. & Lamont, S.J. (1998). Lessons from selection experiments on immune response in the chicken. *Poultry Avian Biology Reviews* 9,125 – 141.

Probabilistic Models of Proteins and Nucleic Acids. *Cambridge University Press*, ISBN 0-521-62971-3

Richard, D., Eddy, S.R., Krogh, A., Mitchison, G. (1998). Biological Sequence Analysis:

Rundicki, W. (2004). Feature synthesis and extraction for the construction of generalized properties of amino acids. *Proc of Rough Sets and Current Trends in Computing: 4th International Conference, Uppsala, Sweden, June 1-5, p. 786 - 791.*

Saitou, N., Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, volume 4, issue 4, pp. 406-425.

Sigrist, C.J, Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., Hulo, N. (2010). "PROSITE, a protein domain database for functional characterization and annotation". *Nucleic Acids Res* 38: 161–166 doi:10.1093/nar/gkp885. PMC 2808866. PMID 19858104.

Sleator, R.D., Walsh P. (2010). "An overview of in silico protein function prediction". *Arch Microbiol* 192: 151–155. doi:10.1007/s00203-010-0549-9. PMID 20127480.

Smith, E.N. and Kruglyak, L. (2008) Gene-environment interaction in yeast gene expression. *PLoS Biol*, 6(4): e83. doi:10.1371/journal.pbio.0060083

Stitziel, N.O., Binkowski, T.A., Tseng, Y.Y., Kasif, S., Liang, J. (2004). topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res*, 32:D520–22.

Stitziel, N.O., Tseng, Y.Y., Pervouchine, D., Goddeau, D., Kasif, S., Liang, J. (2003). Structural location of disease-associated single-nucleotide polymorphisms. *J. Mol. Biol*, 327:1021–30.

Storey, J.D., Akey, J.M., Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol*, 3(8): e267. doi:10.1371/journal.pbio.0030267

Wahlberg, P., Carlborg, Ö., Foglio, M., Tordoir, X., Syvänen, A.-C., et al. (2009). Genetic analysis of an F(2) intercross between two chicken lines divergently selected for body-weight. *BMC Genomics* 10: 248.

Xu, P., Siegle, P.B., Denbow, D.M. (2011) Genetic selection for body weight in chickens has altered responses of the brain's AMPK system to food intake regulation effect of ghrelin, but not obestatin. *Behav. Brain Res,* 221:226.

Yang, M., Li, B., Tomchick, D.R., Machius, M., Rizo, J., Yu, H., Luo, X. (2007). p31comet blocks Mad2 activation through structural mimicry. *Cell*, 131:744-755.

Ye, Y., Godzik, A. (2004). "FATCAT: a web server for flexible structure comparison and structure similarity searching". *Nucleic Acids Res* 32: W582–W585. doi:10.1093/nar/gkh430. PMID 15215455

Acknowledgements

The works of this thesis were performed at Department of Clinical Science, Swedish University of Agriculture Science. The Swedish Foundation for Strategic Research was gratefully acknowledged for financial supports.

Behind this thesis lies effort from many people. I would like to express my sincere gratitude to all of my friends and colleagues help me and support me during that time. My special thanks to:

Stefan Marklund - my main supervisor: this thesis could not have been done without your genuine, generous and cogent supports. You not only served as my supervisor for my academic career, but also become a good friend of mine and encourage me when I was in a plight. Your humor, optimism and kindness give me a bright light in a dark and gloomy tunnel. The small paragraph could not express my gratitude to you. Here, I would like to have my best wish for your new career.

Örjan Carlborg – my co-supervisor: thank you for accommodating me in your group, introducing the world of quantitative genetics and providing experiences comments on my manuscripts. Thanks you for inviting me to the baking party in your house, which is really nice!

Simon Forsberg – Thank you for being nice and trying to challenge me in the gym. P.S. I don't think you could beat me ☺.

Marcin Kierczak – Thank you for collaborating the PASE paper and introducing me the world of amino acid property.

Monika Brandt – Thank you for your kindness and caring. Best wish for your coming Ph.D study!

Muhammad Ahsan – thank you for working together for four years. Your cookies are really delicious!

Xia Shen – Thank you for introducing the information of the group when I was in interview. Good luck for your new job in UK!

Mats Pettersson – Thank you your experienced suggestions on my studies, which really help me a lot!

Zheya Sheng – Thank you for providing amusing news, which really give me a happy mood!

Ronald Nelson – Thank you for making the amazing cover picture of the thesis, and sharing the information and skills of the website.

Yanjun Zan – Thank you for your kindness. Best wish for your future Ph.D study!

To my friends in Stockholm – **Fabio O, Cristina m, Pedro, H, Lena A, Rafiel A, Alain C, Liming B, Allessandra C, Dawei X, Xiaomei D, Elina L, Lennart E, Varinia G, Karina G, Eveline H, Xiang H, Sanna J, Kyriakos K, Magda L, Magnus L, Anqua L, Chunyan L, Lui W, Tara M, Junmei M, Guomin O, Deniz Ö, Lola H, Dongliang Q, Ewa S, KengLing W, JianLiu W, Shen X, Katta F.** I remember the building we live, the birthday party you made for me, the dumplings we made on Chinese new year eve, the community pub we usually sit in, the lake we often have picnic, the firework we watch every new year eve, the big party of Stockholm 750 år. The time with you are awesome, thanks buddies!

To my friends in Gothenburg –**Mari C, Zhen C, Santosh D, Marcela D, Himanshu J, Ying G, Kristin L, Parhati H, Fatima K, Tiange L, Huaqing L, Xue W, Lumi M, Xin L, Ziyu L, Ban W, Zsofia M, Szilard N, Rauan S, Julia S, Si C, Antonio M, David N, Feifei D, Joacim K, Olga R, Alberto C.** When I recall my time in Gothenburg, the first thing coming to my mind is not Chalmers, but weekly afterwork party, plane adventures, basketball matches, pool parties, BBQ in a rainy day,

dancing pub, cooking weekly rotation, conference together in Uemå, and desperately preparing the exam (I hardly forget ☹). You give me a wonderful time in Gothenburg, thanks buddies!

To my friends in Uppsala – **Chrisitina B, Ulrika B, Xiaodong L, Lovisa B, Feifei X, Erik L, Marta S, Da W, Annelie L, Ruixue X, Sofie L, Boyang L, Juliet BL.** I know Uppsala is small and peaceful town. Living here, I may sometimes feel boring but fortunately I meet you guys. I really appreciate the time we spend in downtown cafeteria, cinema, local pub, Italian restaurant, adventure in Stockholm and the new year eve. Thank you for being with me in these four years, all of you are wonderful!