QUANTITATIVE EPIGENETIC ANALYSIS ON A WHOLE GENOME SCALE

by

Yaodong Hu

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Animal Sciences)

at the
UNIVERSITY OF WISCONSIN–MADISON
2015

Date of final oral examination: 05/12/2015

The dissertation is approved by the following members of the Final Oral Committee:

Daniel Gianola, Professor, Animal Sciences
Guilherme J.M. Rosa, Professor, Animal Sciences
Natalia de Leon, Associate Professor, Agronomy
Hasan Khatib, Professor, Animal Sciences
Kent A. Weigel, Professor, Dairy Science

To Olivia, Wei,
and my parents.

Acknowledgments

After I was accepted as a Master's student of Prof. Qin Zhang in the College of Animal Sciences and Technology in China Agricultural University, I had an opportunity to attend the 3rd International Conference on Quantitative Genetics held in Hangzhou, China before I officially started my graduate program. I met many renowned quantitative geneticists in ICQG3, and this experience led me to the decision of pursuing a Ph.D. degree abroad. Luckily, I met Prof. Daniel Gianola in 2008 when he was invited to CAU for a talk on Bayesian prediction of complex traits, and he kindly accepted me as a Ph.D. student in the following year.

I was extremely fortunate to have Prof. Gianola as my major advisor. His scientific attitude had a huge impact on me and he never stopped encouraging me to take good advantage of this research training opportunity and become an independent scientist. I am grateful to my co-advisors Professors Guilherme Rosa and Natalia de Leon for their close guidance on my research projects and their support on my assistantship, especially the Monsanto Fellowship on which I was for three years. I also thank Professors Hasan Khatib and Kent Weigel for their valuable advice on my research. Thank you to all my committee members for your advice, support, and great patience during my long Ph.D. study.

During my Ph.D. career, I received help from many people on either academics or personal life. I would like to take this opportunity to thank Professors Dave Thomas, Brian Kirkpatrick, George Shook and many of my fellow graduate students and post-docs in the Animal Breeding and Genetics group from the Department of Animal Sciences at the University of Wisconsin - Madison. Special thanks go to Rostam Abdollahi-Arpanahi, Huihui Duan, Vivian Felipe, Wen Huang, Nanye

Long, Gota Morota, Francisco Peñagaricano, Paulino Pérez, Bruno Valente, and Chen Yao for their kind help in improving my professional skills. I would also like to thank Dr. Gustavo de los Campos and Dr. Ana Vázquez for serving as excellent graduate student examples, even though we shared only a very short time period after I started my Ph.D. program in the summer of 2009. Further, I thank Prof. Sijian Wang, Dr. Ning Leng, and Dr. Xu Xu from the Department of Statistics for their statistical consulting. Last but not least, I thank Dr. Timothy Beissinger, who was a Ph.D. student in the Department of Agronomy and a fellow recipient of the Monsanto Fellowship, and Dr. Tao Zhang from the Department of Horticulture, for their consulting on plant breeding and methylation quantification regarding part of my research projects.

I also received much help from the faculty of the Department of Animal Sciences. Specifically, I would like to thank Dr. Dan Schaefer and Dr. Ralph Albrecht for providing a graduate assistantship to me when I was in financial difficulty. Most of my work was supported by the Monsanto Fellowship, United States Department of Agriculture Hatch Grant (142-PRJ63CV and 142PRJ28RT), and the Departmental Fund from Animal Sciences.

Finally, I wish to thank all my family members for your support and encouragement during the past several years. I would have not been able to finish my Ph.D. study without your unconditional love.

Table of Contents

List of Tables

List of Figures

Abstract

Epigenetics has attracted increased scientific interests in the last several decades because of its important role on gene expression regulation. Epigenetic modification alters gene function in a mitotically and/or meiotically heritable manner without changing the underlying DNA sequences and, hence, understanding its mechanisms is important in biological research. As a typical epigenetic phenomenon, genomic imprinting results in differential and/or preferential gene expression in a parent-of-origin fashion. Given a potential impact of imprinting on complex traits in both agriculture and epidemiology, investigating the effect of genomic imprinting on the variation of such traits is of interest. According to a recently proposed quantitative genetic model that incorporates imprinting effects by differentiating reciprocal heterozygous genotypes, genomic imprinting contributes to transmissible variation across generations, which hinted at least two possible areas of research relating integrating imprinting into genetic improvement programs. The first one is to investigate its impact on genetic variation in a genome-wide association study (GWAS) context; the second is to assess the usefulness and effectiveness of taking imprinting and/or parent-of-origin effects into account in whole-genome prediction of complex traits. These two aspects, examined with mouse body mass index (BMI) data, constitute two research chapters of this thesis. Single nucleotide polymorphisms (SNP) markers were used as input information in these two studies since the imprinting model employed was adapted from a classical quantitative genetics model that reflects phenotype-genotype associations at the DNA level. However, epigenetic modifications do not change the underlying DNA sequence, indicating that DNA polymorphisms might be inadequate for explaining all epigenetic-induced variation. Therefore, a third study was conducted to evaluate

whether prediction of complex traits can be enhanced using epigenetic polymorphisms, here methylation profiles from a methylated DNA immunoprecipitation (MeDIP) experiment. Prediction models for plant height of *Arabidopsis thaliana* were built non-parametrically via reproducing kernel Hilbert spaces (RKHS) regression. Overall, studies in this thesis tackled phenotype-(epi)genotype associations. The results illustrate the ability of the proposed models of capturing and exploiting epigenetic information, and the importance of epigenetics on the variation of complex traits.

Chapter 1

Introduction

Life sciences aim at a better understanding of biological phenomena. An important subdiscipline of life sciences, genetics, studies genes, heredity, and variations in living organisms. The age of modern genetics began with Gregor Mendel and much accumulation of knowledge has accrued since then.

1.1 A Brief History of Modern Genetics

It is widely accepted that Gregor Mendel was the founder of modern genetics. About one hundred and fifty years ago, Mendel presented his seven-year study on pea hybridization (Mendel, 1866) at two meetings of the Natural Society of Brünn in Moravia (Henig, 2000). In his experiments, Mendel studied several traits that were later discovered to be controlled by a single gene. From his data, he observed that selfing progeny of hybrid F_1 individuals tended to have different phenotypes that segregated at a 3:1 ratio and that segregation was not affected by joint consideration of another trait, which also tended to have a 3:1 segregation ratio. These observations became the famous Mendel's Laws of Segregation and Independent Assortment. Although Mendel did not know what determined the observed segregation, he proposed that each trait was controlled by a pair of "inheritance factors", one coming from each parent. He also speculated that the inheritance

factors could be either "dominant" or "recessive" such that a recessive phenotype is observed only in the absence of a dominant inheritance factor, and the 3:1 ratio can be deduced from simple probabilistic arguments.

Although Mendel's laws are at the core of modern genetics, they did not become widely accepted until rediscovered by Hugo de Vries and Carl Correns in 1900 (Bowler, 2003). In molecular biology, on the other hand, DNA (deoxyribonucleic acid) was extracted from white blood cells by Friedrich Miescher in 1869 (Dahm, 2008), but its function was not known at that time. By 1900, it was known that the chemical components of DNA were phosphate, a sugar, and four heterocyclic bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Shortly after that, the term "gene" was first used in Wilhelm L. Johannsen's book (Johannsen, 1913). However, at that time, it was still unclear what was the carrier of genes and DNA was not connected to the inheritance of traits until 1944 (Avery *et al.*, 1944), when DNA was confirmed as the substance responsible for heredity. Another landmark in genetic research was the work of James Watson and Francis Crick (Watson and Crick, 1953), who shared the Nobel Prize in Physiology or Medicine in 1962 for discovering the 3-D structure of the DNA molecule. Since then, biology experienced a big explosion in the second half of the 20^{th} century. In 1977, with the advent of a DNA sequencing technique invented by Frederick Sanger, another Nobel Prize winner, investigation of variation in life sciences at the DNA level became feasible.

It is now known that in diploid species, each individual carries two copies of each gene, called alleles. Alleles can be interpreted as alternative forms of a gene, some of which are dominant, and some recessive, as described by Mendel. The set of two alleles in an individual is known as the "genotype" at that locus. As an illustration, consider human's blood types as described by Landsteiner (Landsteiner, 1900). Four possible blood types are A, B, AB, and O, determined by different combination of three alleles: A, B, and O, where A and B are dominant alleles and O is recessive. Thus, both genotypes AA and AO produce blood type A, and the same is true for blood type B. However, blood type AB is observed when the genotype is AB and blood type O is

observed if and only if the genotype is OO. This simple example shows the relationship between alleles, genotypes, and phenotypes in a categorical (or discrete) trait, which is not affected by environmental factors. In a continuous trait like human height, there are usually multiple small-effect gene loci that affect the phenotype jointly, together with an environmental component. Such traits are generally modeled using the infinitesimal model proposed by Ronald A. Fisher and Sewall G. Wright (Fisher, 1918; Wright, 1921). In this case, the trait is referred to as "complex trait" and alleles at each contributing locus are not simply distinguished as dominant or recessive. Rather, different alleles have a different "effect size" on the phenotype, and these can be inferred using quantitative genetic models. Regardless of whether a trait is discrete or complex, alleles at a single locus are viewed as segregating according to Mendel's first law. When the investigation involves two or more loci (or genes), as in the case of complex traits, Mendel's Law of Independent Assortment or concepts based on Morgan's ideas on crossing-over and linkage (see below) apply.

Linkage was first observed by Thomas H. Morgan in 1910 in *Drosophila melanogaster* (the fruit fly). He found that the inheritance of eye color (red or white) tended to be associated/linked with the sex of an individual, a phenomenon he termed sex-linked inheritance (Morgan, 1910). Later, his student Alfred Sturtevant constructed the first genetic map (Sturtevant, 1913) and claimed that genes were linearly located on chromosomes. Morgan and Sturtevant hypothesized that genes located on the same chromosome would be linked together during meiosis. The degree of linkage was mostly dependent on the physical distance between these genes, and he posed that linked genes might be subject to crossing-over within a pair of homologous chromosomes, which can lead to recombination of genotypes at different gene loci. Linkage (and the probability of crossing-over) is extremely useful in genetics, providing, for example, information for gene mapping. From 1920's to 1930's, as knowledge accumulated, it became widely accepted that the gene was the functional unit for phenotypes, and finding the locations of genes on chromosomes became of interest. The foundations of gene mapping were subsequently laid by John B. S. Haldane, Lancelot Hogben, R. A. Fisher, Lionel Penrose, and Newton E. Morton (Morton, 1955), among others. If a segment of

DNA with known location and known to have no effect on phenotype was used as a flag, then there could be a functional gene locus linked to this marker (usually expected to be in close vicinity of the marker locus) such that the segregation at the marker locus was associated with variation in the phenotype of interest. Although the rationale of gene mapping was simple, linkage analysis in humans, animals and plants languished for more than a half century after Sturtevant's discovery because of the lack of genetic markers and of adequate computing power. Mapping efforts were speeded up by the development of DNA markers that provided a virtually unlimited supply of genetic markers – an idea first conceived by Botstein and colleagues for yeast crosses (Petes and Botstein, 1977) and subsequently for human families (Botstein *et al.*, 1980). A decade after the discovery of DNA markers, the first genome map in plants was reported in maize and tomato (Helentjaris *et al.*, 1986) using RFLP (restriction fragment length polymorphism) markers in an F_2 population. Since then, genetic markers have dominated much of recent genetic studies, mainly because they are much easier to observe than the genes themselves.

With the polymerase chain reaction (PCR), invented by Kary Mullis in 1983 (Mullis, 1990), only a small amount of DNA was necessary for sequencing, and the genomic era of biology emerged. Nowadays, markers at single nucleotide resolution (the so-called single nucleotide polymorphism, SNP) are widely available on major species and have become one of the most important tools in genetic studies. With the launch of the Human Genome Project in 1990, there has been an unprecedented explosion of biological information, which made biology and genetics one of the fastest-developing sciences. Because of this, the 21st century is usually referred to as the century of biology (Venter and Cohen, 2004).

Today, with the integration of information technology, genetics is influencing every aspect of the life sciences. In humans, genetic information can be used to map disease genes, as anticipated by David Botstein and colleagues in 1980. Ten years later, the first approved gene therapy on humans was used by William French Anderson to treat a 4-year-old girl suffering from SCID (severe combined immunodeficiency). Subsequently, more disease-related genes have been discovered, e.g.,

the *HTT* gene responsible for Huntington's disease was isolated by the US - Venezuela Huntington's Disease Collaborative Research Project in 1993; the *BRCA1* gene responsible for breast cancer was discovered in 1994, and gene therapies have been developed accordingly. Besides its application for finding causal mutations, genomic information can also be used to predict the susceptibility to diseases (Wray *et al.*, 2007, 2008; de los Campos *et al.*, 2010). Since complex diseases are often the consequence of both genetic and environmental factors, people with high disease risks can be advised to modify their life style, which may reduce the probability of disease onset later in life. This is an application of "personalized medicine", which is widely believed to be the future of health care.

In addition to human epidemiology, genetics is also playing a big role in many other scientific realms, one of which is agriculture. Artificial selection of plants and animals represents an application of genetics to agriculture, and has contributed enormously to the improvement of production, reproduction, and quality traits of agricultural important species.

1.2 Artificial Selection – An Agricultural Application of Genetics

Charles Darwin, who developed the theory of natural selection in his *On the Origin of Spieces* (Darwin, 1859), pointed out that mutation was the source of variation, and that only those individuals with the highest fitness relative to some environment would survive from the selection by nature. In agriculture, selection has existed for thousands of years as well, but the selection pressure comes from humans, so it is called artificial selection. Artificial selection is probably the most important tool to adapt agricultural species in the direction of human demands. Although modern breeding is now very precise, it used to be very empirical in the earlier times: keeping individuals with the best appearance and discarding the worst ones was no doubt the earliest application of artificial selection, even before Darwinism. Later, empirical assortative mating and inbreeding led to the concept of "breed". For example, Robert Bakewell, a British agriculturalist, bred Dishley Longhorn

beef cattle, Leicester sheep, and Shire horses in the 18th century using selective breeding. His work not only led to specific improvements in major domestic livestock, but also contributed significantly to the general knowledge of artificial selection, which resulted in the formation of ten cattle breeds, twenty swine breeds, six horse breeds, and more than thirty sheep breeds in UK during 100 years. Due to his outstanding accomplishments, he is often acknowledged as a pioneer of modern animal breeding (Wood and Orel, 2001).

When modern genetics became appreciated by the scientific community in the early 20th century, animal and plant breeding started to make use of knowledge from disciplines like statistics, and was transformed from an art into a science. The first application of statistics on agriculture was probably Gregor Mendel's plant hybridization experiment that led to the famous Law of Segregation, but R. A. Fisher was widely considered as the founder of biometrics and quantitative genetics, in which statistical methods were intensively involved. In 1918, Fisher introduced the concept of additive relationships to describe kinship between relatives (Fisher, 1918), and the infinitesimal model he proposed became a standard in the study of complex traits for about 100 years. Later, based on his study on crop variation in 1920's, Fisher introduced the analysis of variance (ANOVA), a method that can be used to partition the observed variance into components due to experimental design or environmental effects (Fisher, 1921–1924). When the experimental factors are genetic, the total variance is partitioned into genetic and environmental components. In addition to variance decomposition, integrating the work of Fisher and Wright and the population genetics theory of Godfrey H. Hardy (Hardy, 1908) and Wilhelm Weinberg (Weinberg, 1908) formed the basic pillars of quantitative genetics. Based on these theories, a breeding program would be most effective if candidates were selected by their genetic merits/breeding values, a metric describing the expected change of population mean if an individual was used for reproduction.

Quantitative genetics accelerated animal and plant breeding significantly. However, breeders did not fully benefit from the theory at the onset, probably because: 1) early statistical models did not predict breeding values with enough accuracy and, 2) lack of computational tools limited the

use of the theory. This situation gradually changed, and Charles R. Henderson proposed the best linear unbiased predictor (BLUP) shortly after he obtained his Ph.D. degree (Henderson, 1975, 1984) at Iowa State University. In BLUP, both fixed effects and random effects of a mixed model can be inferred simultaneously, so systematic effects can be corrected while predicting the breeding values, part of the random effects. By doing this, more accurate estimates and predictions can be obtained. In addition, breeding values of individuals without observations can be predicted from relatives with production records, using the additive relationship matrix derived from a pedigree. Breeding values of a large number of individuals can then be obtained from measurements on a smaller group of individuals and from the lineage. Starting in the 1980's, powerful computer programs became available for solving large equation models, so BLUP became widely used in breeding programs around the world and has dominated animal breeding for several decades.

Indeed, statistics has been aiding breeding programs for over 100 years (Gianola and Rosa, 2015). The contribution of molecular genetics to artificial selection must be acknowledged as well. After about a century of research, it has been recognized that although many economically important traits in agriculture are controlled by many genes, not all genes have small effects, as represented in the infinitesimal model. Instead, some traits (e.g., meat quality in pigs) are known to be controlled by few large-effect genes (major genes) in addition to a large number of small-effect genes, which is known as the major gene(s) model (Hayes and Goddard, 2001). Estimating the effect of major genes on phenotype and finding the locations of such genes has always been of interest, and these studies would not be possible without the use of genetic markers. In early stages, a "major gene" was viewed as a block of DNA segment, termed as quantitative trait loci (QTL), perhaps containing several genes. Early QTL mapping models often assumed that there was only one large effect QTL in the genome, i.e., the single QTL model (Lander and Botstein, 1989; Haley and Knott, 1992). The single QTL model was useful for finding QTL with very large effects on the trait of interest, but missed moderate or small effect QTLs. Due to the advent of more dense marker panels, later studies were able to assume multiple QTLs throughout the genome, and this development was

helpful for the study of polygenic systems (e.g., Meuwissen and Goddard, 2004).

Despite the current ability of locating major genes, animal and plant breeding programs still need to be carefully designed to be able to efficiently use such information. On one hand, for traits where major genes operate, keeping individuals with favorable alleles is a widely used strategy. This procedure is called Marker Assisted Selection (MAS, Ribaut and Hoisington, 1998), which is still popular in plant breeding (e.g., Collard and Mackill, 2008). In some cases, if the favorable allele is not present in the current breeding material, it can be introduced by either crossing with another population or by gene modification, known as transgenic or gene modification (GM) technology. On the other hand, although QTL mapping can assist artificial selection, it is also true that the large amount of small-effect loci cannot be located via QTL mapping and applied in MAS. Even with high density SNP markers and with genome-wide association studies (GWAS, Hindorff *et al.*, 2014), considered as a high-resolution counterpart of QTL mapping, many small-effect SNP markers are still missed due to the lack of statistical power. Further, there are many complex traits for which the infinitesimal model provides a better approximation than a major gene model, in which case QTL mapping and GWAS would likely fail (MacArthur, 2008). In this case, prediction of breeding values remains as a preferred method for artificial selection.

Breeding values can be predicted using phenotypic and pedigree records through use of mixed model methodology (e.g., BLUP, Henderson, 1984; Mrode, 2014). However, as high-density SNP marker chips covering the whole genome of major species became available at a relatively low cost in recent years, it became possible to predict genetic merit using genomic information, the so-called genomic estimated breeding value (gEBV); selection according to this gEBV is known as genomic selection (GS, Meuwissen *et al.*, 2001). Genome-enabled predicted breeding values are expected to be more accurate than those obtained from pedigree-based BLUP since, in the latter, the relationship between relatives is an "expected" measurement of kinship that relies on the proportion of identical-by-descent (IBD) alleles shared by related individuals, whereas in GS genomic information at the DNA level provides a realized kinship, based on a similarity in state.

Another advantage of GS is that genomic information can be obtained at a very young age of breeding candidates, much earlier than when a record of performance or progeny test information becomes available. This is especially useful in dairy cattle breeding and can reduce the generation interval and breeding costs significantly (Schaeffer, 2006). Also, GS is useful for traits that are hard to measure, for example in carcass traits where slaughter of animals is required, or traits that are measured late in life, such as longevity. Because of its attractive features, GS became very popular in animal and plant breeding in recent years.

1.3 Beyond the Central Dogma of Biology

In molecular biology, the central dogma proposed by F. Crick (Crick, 1956, 1970) provides an explanation of the flow of genetic information within a biological system. The central dogma states that genes, as functional segments of DNA molecules, are transcribed into RNAs, and RNAs are translated into proteins according to their sequence and a 3-mer genetic code called codon, which determines the order of amino acids in proteins. Because the biological pathway points from DNA to protein, it is usually believed that variation in a DNA sequence correlates with changes in phenotype, and this is the basis of associating phenotypic variation through changes in genomic information. However, the same gene can be expressed differently in different tissues or in different stages of life. If these spatial and temporal differences are to be considered, the classical central dogma should be supplemented by information at the level of gene expression regulation, which is one main topic of epigenetic studies.

Epigenetic gene regulation has been studied for years, especially in human genetics, for the reason that it is associated with some genetic diseases (e.g., Tollefsbol, 2012). In agriculture, on the other hand, it has not recieved much attention until recently. This thesis reports results of some studies on epigenetics, of its relationship with animal/plant breeding and of its potential application. This thesis is organized as follows: Chapter 2 reviews concepts of epigenetics and

genomic imprinting, a typical epigenetic phenomenon. Some relationships with human diseases and agricultural production are discussed as well. Chapter 3 discusses a quantitative genetic model incorporating imprinting and its application in some recent studies. This imprinting model defines half the contrast between the genotypic values of the two reciprocal heterozygotes as an imprinting effect (i), in addition to an additive effect (a) and a dominance effect (d) in a standard quantitative genetic setting. Chapters 4 and 5 include two studies using the imprinting model introduced in Chapter 3. In Chapter 4, the imprinting model was employed to conduct a GWAS-like study on mouse body mass index (BMI) using SNP markers as genomic proxies. Results indicate that the additive genetic variation would be underestimated if an additive model was applied on a trait affected by an imprinting-induced parent-of-origin effect (POE). In Chapter 5, the same model was extended to the entire genome to carry out a whole-genome prediction of mouse BMI. The main purpose of this study was to evaluate whether including POE can improve predictive performance when it is present. No advantage of incorporating POE was found in this particular data set.

Chapter 6 describes a non-parametric prediction study using epigenetic information. As epigenetic modification changes gene expression without modifying the underlying DNA sequence, epigenetic information might enhance phenotypic prediction if variation is observed at the epigenome level. This study used isogenic *Arabidopsis thaliana* epigenetic recombinant inbred lines (epiRILs) as material and predicted plant height using DNA methylation data obtained from a methylated DNA immunoprecipitation (MeDIP) experiment. Results suggested that epigenetic information was useful for whole-genome prediction, and a higher predictive ability was expected when supplemented to genomic polymorphisms at the DNA level.

Finally, Chapter 7 offers an overall discussion and summary of the thesis. Some future perspectives on applications of epigenetics in agriculture and human genetics are given in this chapter as well.

References

Avery, O. T., C. M. Macleod, and M. McCarty, 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. J. Exp. Med., 79(2): 137–158

Botstein, D., R. L. White, M. Skolnick, *et al.*, 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet., 32(3): 314–331

Bowler, P., 2003. *Evolution: The History of an Idea.* University of California Press. ISBN 9780520236936

Collard, B. C. and D. J. Mackill, 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc. Lond., B, Biol. Sci., 363(1491): 557–572

Crick, F., 1956. On protein synthesis. Symp. Soc. Exp. Biol. XII, pp. 139–163 (early draft of original article)

Crick, F., 1970. Central dogma of molecular biology. Nature, 227(5258): 561–563

Dahm, R., 2008. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum. Genet., 122(6): 565–581

Darwin, C., 1859. *On the Origin of Species: By Means of Natural Selection.* John Murray

de los Campos, G., D. Gianola, and D. B. Allison, 2010. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet., 11(12): 880–886

Fisher, R. A., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb., 52: 399–433

Fisher, R. A., 1921–1924. Studies in crop variation. I – III

Gianola, D. and G. J. M. Rosa, 2015. One hundred years of statistical developments in animal breeding. Annu. Rev. Anim. Biosci., 3: 19–56

Haley, C. S. and S. A. Knott, 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity (Edinb), 69(4): 315–324

Hardy, G. H., 1908. Mendelian proportions in a mixed population. Science, 28(706): 49–50

Hayes, B. and M. E. Goddard, 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol., 33(3): 209–229

Helentjaris, T., M. Slocum, S. Wright, *et al.*, 1986. Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor. Appl. Genet., 72(6): 761–769

Henderson, C. R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, 31(2): 423–447

Henderson, C. R., 1984. *Applications of Linear Models in Animal Breeding.* University of Guelph, Ontario, Canada

Henig, R., 2000. *The Monk in the Garden: The Lost and Found Genius of Gregor Mendel, the Father of Genetics.* Houghton Mifflin

Hindorff, L. A., J. MacArthur, J. Morales, *et al.*, 2014. A catalog of published genome-wide association studies. http://www.genome.gov/gwastudies/

Johannsen, W. L., 1913. *Elemente der exakten Erblichkeitslehre: mit Grundzügen der biologischen Variationsstatistik.* Gustav Fischer

Lander, E. S. and D. Botstein, 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121(1): 185–199

Landsteiner, K., 1900. Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. Centralblatt für Bakteriologie, Parasitenkunde und Infektionskrankheiten, 27: 357–362

MacArthur, D., 2008. Why do genome-wide scans fail? Genetic Future (online), http://www.wired.com/2008/09/why-do-genome-wide-scans-fail/

Mendel, G., 1866. Experiments in Plant Hybridization (Translated from original German version: Versuche über Pflanzen-Hybriden). Proceedings of the Natural History Society of Brünn

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4): 1819–1829

Meuwissen, T. H. E. and M. E. Goddard, 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol., 36(3): 261–279

Morgan, T. H., 1910. Sex limited inheritance in *Drosophila*. Science, 32(812): 120–122

Morton, N. E., 1955. Sequential tests for the detection of linkage. Am. J. Hum. Genet., 7(3): 277–318

Mrode, R., 2014. *Linear Models for the Prediction of Animal Breeding Values.* CAB International, 3rd edition

Mullis, K. B., 1990. The unusual origin of the polymerase chain reaction. Sci. Am., 262(4): 56–61

Petes, T. D. and D. Botstein, 1977. Simple Mendelian inheritance of the reiterated ribosomal DNA of yeast. Proc. Natl. Acad. Sci. U.S.A., 74(11): 5091–5095

Ribaut, J.-M. and D. A. Hoisington, 1998. Marker-assisted selection: new tools and strategies. Trends Plant Sci., 3: 236–239

Schaeffer, L. R., 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet., 123(4): 218–223

Sturtevant, A. H., 1913. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. J. Exp. Biol., 14: 43–59

Tollefsbol, T. (Editor), 2012. *Epigenetics in Human Disease.* Academic Press, MA, USA

Venter, C. and D. Cohen, 2004. The Century of Biology. New Perspectives Quarterly, 21(4): 73–77

Watson, J. D. and F. H. Crick, 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature, 171(4356): 737–738

Weinberg, W., 1908. Über den Nachweis der Vererbung beim Menschen. Jahreshefte Verein f. vaterl. Naturk. in Württemberg, 64: 368–382

Wood, R. and V. Orel, 2001. *Genetic Prehistory in Selective Breeding: A Prelude to Mendel.* Oxford University Press

Wray, N. R., M. E. Goddard, and P. M. Visscher, 2007. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res., 17(10): 1520–1528

Wray, N. R., M. E. Goddard, and P. M. Visscher, 2008. Prediction of individual genetic risk of complex disease. Curr. Opin. Genet. Dev., 18(3): 257–263

Wright, S., 1921. Systems of mating. Parts I – V. Genetics, 6: 111–178

Chapter 2

Epigenetics and Genomic Imprinting

Proteins are the building blocks of all living organisms; they are coded by different genes, which range from several hundreds in *Nanoarchaeum equitans* to tens of thousands in some plant species. Genes locate linearly in the genome and the central dogma suggested that the genes are basis that determines phenotypes. However, this classical view does not consider epigenetic regulation of gene expression that, biologically, is in between the DNA and the RNA layers. In this chapter, concept of epigenetics and genomic imprinting are introduced and known mechanisms of imprinting are briefly discussed.

2.1 Epigenetics

First introduced by Conrad H. Waddington around 1940 to describe "the interactions of genes with their environment, which bring the phenotype into being" (Waddington, 1939, 2012), the term "epigenetics" has changed its meaning in the past fifty years. Although various definitions of this term have been offered (e.g., Jablonka and Lamb, 2002; Holliday, 2006; Bird, 2007; Ptashne, 2007; Krause *et al.*, 2009), a current consensus definition of "epigenetics" is that it is the study of mitotically and/or meiotically heritable variations in gene function that do not involve changes

in DNA sequence (Riggs *et al.*, 1996; Riggs and Porter, 1996). Epigenetic regulation can manifest as commonly as the manner in which cells terminally differentiate into skin cells, liver cells, brain cells, for example. But also, it can have more damaging effects that will result in diseases like cancer (e.g., Jones and Baylin, 2007; Esteller, 2008; Virani *et al.*, 2012), if dysregulated. Epigenetic regulation results from various epigenetic modifications, among which DNA methylation and histone modifications are two major forms (Figure 2.1; Kouzarides, 2007; Rivera and Bennett, 2010; Moore *et al.*, 2013). In recent years, non-coding RNAs (ncRNAs) were found to represent a hidden layer of internal signals that control various levels of gene expression associated with physiological and developmental processes. Their role in epigenetic regulation of gene expression has been acknowledged as well (Zhou *et al.*, 2010; Kaikkonen *et al.*, 2011).

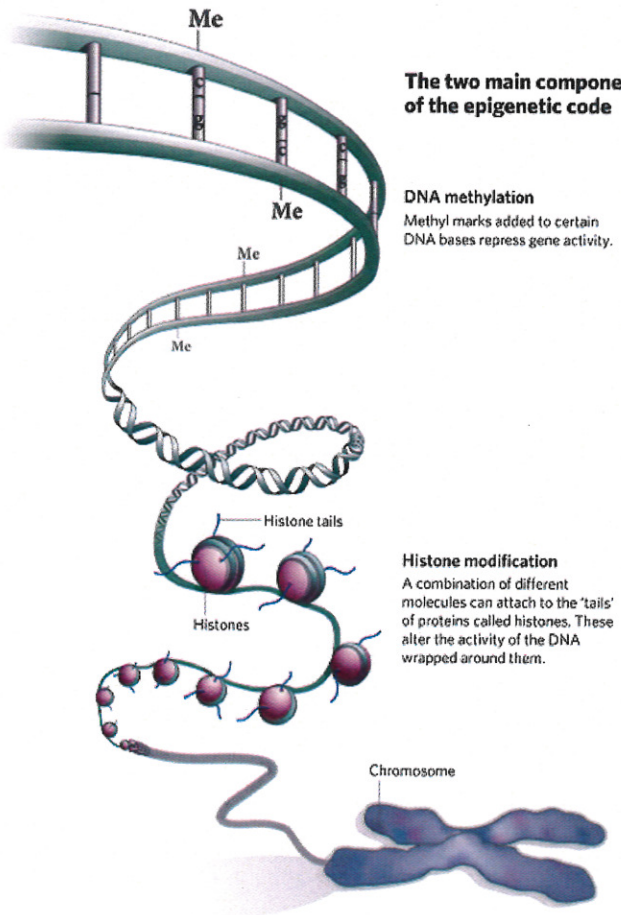
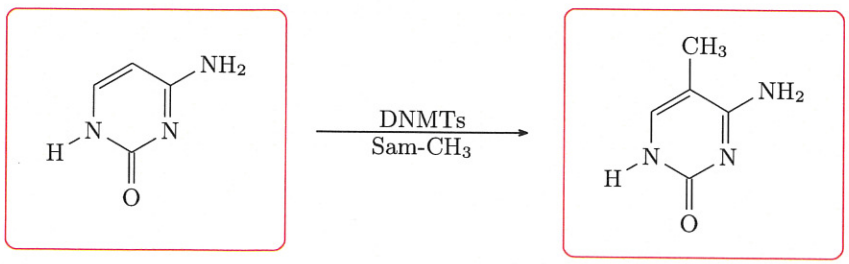


Figure 2.1: The two major forms of epigenetic modification (Qiu, 2006): DNA methylation (top) and histone modification (bottom).

2.1.1 DNA Methylation and CpG Islands

DNA methylation is the first recognized and most well-characterized epigenetic modification. It is the covalent addition of a methyl group (-Ch_3) to a nucleobase. Up to date, the known targets of DNA methylation are cytosine and adenine, resulting in methylcytosine and methyladenine. It is also known that the methyl group can be added to either the 5-position carbon atom (C_5) or the 4-position nitrogen atom (N_4) of the cytosine pyrimidine ring, resulting in 5-methylcytosine (^{m5}C) or 4-methylcytosine (^{m4}C), respectively; or it can be added to the 6-position nitrogen atom (N_6) of the adenine purine ring, forming 6-methyladenine (^{m6}A) (Ratel *et al.*, 2006). It has been found that ^{m4}C is only encountered in bacterial DNA (Ehrlich *et al.*, 1987), and ^{m6}A is mainly found in mitochondrial DNA of flowering plants (Vanyushin, 2006), such that ^{m5}C is widely believed to be the sole form of DNA methylation in nuclear DNA in vertebrates and higher plants (Jeltsch, 2002; Meissner *et al.*, 2005). Therefore, only ^{m5}C is discussed here. In ^{m5}C (Figures 2.2, 2.3), S-adenosyl-L-methionine (SAM-CH_3) is acting as a methyl group donor and the methyl group is transfered from SAM-CH_3 to cytosine under the control of a family of enzymes called DNA methyltransferases (DNMTs), including DNMT1, DNMT3a and DNMT3b. DNMT3a and DNMT3b are *de novo* methyltransferases, preferentially targeting unmethylated cytosine to mediate establishment of new methylation, whereas DNMT1 plays an important role in maintaining methylation status (Bestor and Verdine, 1994; Cheng, 1995; Pfeifer, 2000).



Cytosine 5-Methylcytosine

Figure 2.2: Methylation of Cytosine. A methyl group is added to the 5-position carbon atom.

In mammalian genomes, methylation occurs mainly at cytosines in a CpG dinucleotide context

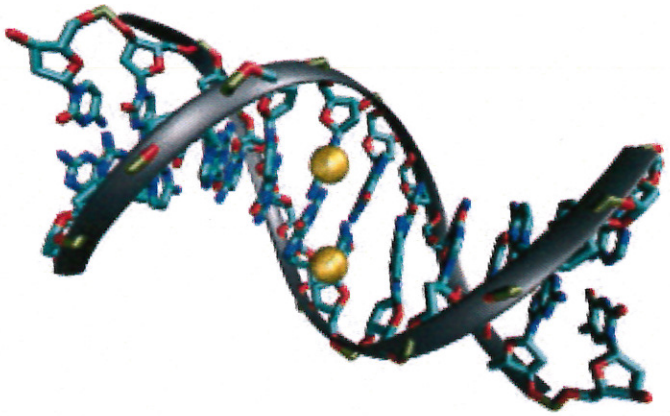


Figure 2.3: DNA (cytosine) methylation (Bock and Lengauer, 2008). Two big golden circles represent ^{m5}C, with methyl groups projecting into the major groove of DNA.

(cytosine followed by guanine, "p" indicates the phosphate bond in between). In plants, on the other hand, ^{m5}C can also occur in CpHpG or CpHpH contexts, where H can be adenine, cytosine, or thymine (Lister *et al.*, 2008). Recently, studies suggested that CpHpG and CpHpH methylation patterns were observed in mammalian genomes as well (Lister *et al.*, 2009; Yan *et al.*, 2011; Xie *et al.*, 2012; Shirane *et al.*, 2013). In humans, approximately 70 $\sim$ 80% CpG sites are methylated (Chen and Riggs, 2011), covering approximately 1.5% of genomic DNA (Lister *et al.*, 2009), and such sites tend to locate in genomic regions sparse in CpG contents. Regions with high CpG densities are usually unmethylated, which are called CpG islands (CGI, Gardiner-Garden and Frommer 1987). CGIs, accounting for $\sim$ 7% of all genomic CpG sites (Bell, 2013), are often found in gene promoter regions (e.g., they overlap the promoter regions of 60 $\sim$ 70% of all human genes, Lander *et al.* 2001; Weber *et al.* 2007) and are characterized by their length ($>$ 200 bp), CG content ($>$ 50%), and observed-to-expected CpG ratio ($>$ 0.6) (Gardiner-Garden and Frommer, 1987). Despite the high density of repeated CpG sequence inside CGIs, the CpG dinucleotide is not rich across the whole genome. In human genome for example, CpG content is only 21% of what was expected (Lander *et al.*, 2001). This phenomenon is called "under representation" of CpG dinucleotides. One explanation for this observation is that methylated CpG sites have a high spontaneous deamination rate that converts ^{m5}CpG into TpG and into CpA on the complementary

strand (Illingworth and Bird, 2009; Matsuo *et al.*, 1993), resulting in a lower frequency of CpG dinucleotides than expected.

2.1.2 Histone Modification

DNA is the genetic material that contains the instructions needed for normal development and functioning of almost all living organisms. However, the linear length of naked DNA far exceeds the microscopic dimensions of a cell nucleus. Hence, in order to fit DNA within the confines of a nucleus, genomic DNA in eukaryotic cells is packaged with special proteins termed histones to form protein-DNA complexes, the chromatin. The basic unit of chromatin is the nucleosome, which is composed of $\sim$ 146 base pairs (bp) of DNA wrapped twice around an octamer of the four core histones: histone 2A (H2A), histone 2B (H2B), histone 3 (H3) and histone 4 (H4) (Luger *et al.*, 1997). Another histone, termed linker histone 1 (H1), interacts with DNA links between nucleosomes and functions in compacting chromatin into higher-order structures that comprise chromosomes. This organization of chromatin allows DNA to be tightly packaged, accurately replicated, and sorted into daughter cells during mitosis (Groth *et al.*, 2007; Routh *et al.*, 2008).

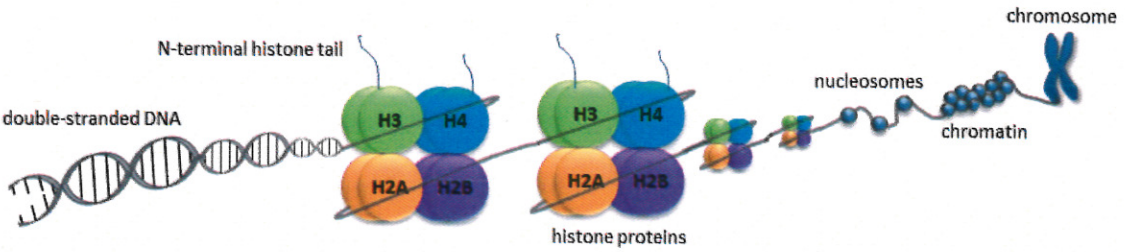


Figure 2.4: Structure of chromosome, chromatin, nucleosome, histone, and DNA molecule shown by schematic representation depicting the organization and packaging of genetic material (http://www.whatisepigenetics.com/histone-modifications/). Nucleosomes are represented by DNA (grey) wrapped around eight histone proteins, H2A, H2B, H3, and H4 (colored circles). N-terminal tails (blue) are shown protruding from H3 and H4.

The core histones are tightly packed in globular regions except for an unstructured part of the polypeptied chain (N-terminal tail) that extends from the globular region (Figure 2.4). The protruding N-terminals of the core histones are subjected to several types of multivalent modifications,

including acetylation, methylation, phosphorylation, ubiquitination, sumoylation, etc (Kouzarides, 2007; Ruthenburg *et al.*, 2007). Histone modifications, recognized as a post-translational modification (PTM), are critical for regulating chromatin structure and function, which can in turn affect many DNA-related processes, such as transcription, recombination, DNA repair and replication, and chromosomal organization. Acetylation/deacetylation and methylation/demethylation of lysines at histone tails are the two most common PTMs in both euchromatin and heterochromatin, the transcriptionally active or inactive states of chromatin, respectively (Jenuwein and Allis, 2001; Kouzarides, 2007; The ENCODE Project Consortium, 2007).

Histone acetylation occurs by the enzymatic addition of an acetyl group (-$COCH_3$) to the lysine residues at the N-terminal of histone tails from acetyl coenzyme A (Acetyl-CoA), mediated by histone acetyltransferases (HATs). Conversely, acetylated histones can be deacetylized by the hydrolytic removal of acetyl groups from histone lysine residues. This process is catalyzed by histone deacetylaces (HDACs) and the removed acetyl group is transferred to coenzyme A (Kuo and Allis, 1998). Histone acetylation generally creates an accessible chromatin conformation while histone deacetylation, often coupled with histone methylation, initiates a compressed chromatin structure that promotes silencing and the formation of heterochromatin (Berger, 2002). Histone methylation, on the other hand, is the transfer of one (monomethylation), two (dimethylation), or three (trimethylation) methyl groups from SAM-CH_3 to lysine or arginine residues of histone proteins by histone methyltransferases (HMTs) (Kouzarides, 2007). However, unlike histone acetylation that mostly results in an enhanced expression, histone methylation can confer both an active or repressed transcriptional state depending on which lysine is methylated (Cheung and Lau, 2005). Analogous to deacetylation, methyl group can be removed from methylated histones as well by histone demethylases (Cheung and Lau, 2005).

2.1.3 Non-coding RNA

Recent high-throughput transcriptomic analyses revealed that eukaryotic genomes transcribe up to 90% of the genomic DNA (The ENCODE Project Consortium, 2004). However, only $1 \sim 2\%$ of these transcripts encode for proteins (Kaikkonen *et al.*, 2011), whereas the vast majority are not translated into proteins, known as non-coding RNAs (ncRNAs). ncRNAs can be divided into infrastructural ncRNAs and regulatory ncRNAs. All epigenetic-related ncRNAs are included in the regulatory ncRNA class and can be divided into two main groups by their length: short ncRNAs that contain microRNAs (miRNAs, $20 \sim 24$ nt; "nt" stands for nucleotide), small interfering RNAs (siRNAs, $20 \sim 24$ nt), piwi-interacting RNAs (piRNAs, $24 \sim 31$ nt) and long ncRNAs (lncRNAs, >200 nt) (Ponting *et al.*, 2009). Regulatory ncRNAs appear to comprise a hidden layer of internal signals that control various levels of gene expression associated with physiological and developmental processes. Their role in epigenetic regulation of gene expression has been reviewed elsewhere (e.g., Zhou *et al.*, 2010; Kaikkonen *et al.*, 2011) and hence is not highlighted here.

2.1.4 Epigenetically Interfered Gene Expression Regulation

DNA methylation has a strong impact on gene expression. Although it does not alter the way in which DNA is transcribed into mRNA (Robertson, 2005; Bock and Lengauer, 2008), it is linked to transcription and gene expression in at least two mechanisms: 1) it directly fosters a locally more compact chromatin structure and hence inhibits the binding of specific transcription factors (Kim *et al.*, 2003; Strunnikova *et al.*, 2005); 2) it attracts methyl-CpG-binding domain proteins which will bring a repressor to silence transcription indirectly (Bird and Wolffe, 1999). Thus, DNA methylation is highly associated with reduced gene expression and is usually regarded as the most important epigenetic mechanism (Bird, 1984; Razin and Cedar, 1991). Mechanism of expression regulation related to histone modification involves accessibility of transcription factors as well, similar to that of DNA methylation (Jones *et al.*, 1998; Nan *et al.*, 1998; Bird and Wolffe, 1999).

Epigenetically interfered gene expression can have various forms. One important aspect is differential gene expression (DE). DE can be reflected in both spatial (the "where") and temporal (the "when") manner, and examples are provided next. Spatially, a gene controlling the accumulation of melanin on skin will not be expressed in blood. Temporally, a gene controlling menstruation (by changing the hormone level) may have an altered behavior after adolescence. Therefore, well regulated expression behaviors are very important for normal development and a dysregulation may result in severe disorders including cancers (Jiang *et al.*, 2004; Tollefsbol, 2012; Pembrey, 2012; Murrell *et al.*, 2013). Another important aspect of epigenetic regulation is allelic inactivation. It is known that all bi-sexual organisms have two sources of genome contributions, one coming from the father and the other coming from the mother. When a gene is under a Mendelian inheritance mode, the genetic contribution from father and mother are equal. However, this is not the case for sex linked genes. Humans, for example, have heterogametic karyotype XY in males and homogametic karyotype XX in females. Thus, for any genes located on the X-chromosome, females have two copies and males have only one copy. In order to compensate the gene dosage, one of the two X-chromosomes in female is epigenetically inactivated. This process, called X-chromosome inactivation (XCI), was first observed in 1959 (Ohno *et al.*, 1959). In plants, mainly observed in snapdragon and maize, a process called paramutation is often associated with gene inactivation as well. Paramutation, induced by interactions between related genes, is the directed heritable change of one allele at a locus being exposed to another one in heterozygote (Brink *et al.*, 1968). Recent study suggested that paramutation is caused by RNA-directed DNA methylation and will lead to change of expression or even inactivation of paramutant, the epigenetically altered allele (Alleman *et al.*, 2006).

2.2 Genomic Imprinting

2.2.1 Definition and Observations

Besides XCI and paramutation introduced above, a third type of gene inactivation called genomic (or genetic, gametic) imprinting is commonly observed in mammals and flowering plants. Imprinting is a special case of epigenetic regulation such that the expression of one copy of a same gene is repressed when this gene copy is inherited from a specific parent, i.e., it is a preferential expression of certain genes depending on whether the genetic material has been inherited from the mother or the father. If a gene is expressed only when it was inherited from the father, it is called maternally imprinted (or paternally expressed) gene, e.g., the *Igf2* (insulin like growth factor II) gene (DeChiara *et al.*, 1991); if a gene is expressed only when it was inherited from the mother, it is called paternally imprinted (or maternally expressed) gene, e.g., the *Igf2r* (*Igf2* receptor) gene (Barlow *et al.*, 1991).

Genomic imprinting in mammals was first discovered in mice in uniparental disomy (UPD) experiments showing that both paternal and maternal genomes are required for a normal fetal development (McGrath and Solter, 1984; Barton *et al.*, 1984). Shortly after that, tens of imprinted genes have been identified in various eutherian species including human, rat, sheep, swine, cattle, etc (Morison *et al.*, 2005) and angiosperms (Matzke and Matzke, 1993; Feil and Berger, 2007). In marsupials (O'Neill *et al.*, 2000; Suzuki *et al.*, 2005; Renfree *et al.*, 2008) and some species of insects (Lloyd, 2000; Khosla *et al.*, 2006; Anaka *et al.*, 2009), there are evidences for the presence of imprinted genes as well. Even in fish (McGowan and Martin, 1997) and nematodes (Bean *et al.*, 2004; Sha and Fire, 2005), imprinted chromatin regions were found in transgenic individuals, although not identified endogenously. Up to now, around 200 mammlian imprinted genes have been documented (`http://igc.otago.ac.nz/home.html`, last updated in Jan, 2011). Comprehensive studies in plants and animals suggested that approximately 2% of all genes are imprinted (Gehring, 2013). Although more imprinted genes have been reported either by prediction (Luedi *et al.*, 2005)

or transcriptome sequencing experiment (Gregg *et al.*, 2010b), there is currently no consensus on the number of imprinted genes in the mammalian genome (Kelsey and Bartolomei, 2012).

At its first discovery, genomic imprinting was believed to play an important role in normal fetal development (McGrath and Solter, 1984; Barton *et al.*, 1984) by observing that individuals with duplicated genomes from a single parent (either maternal, gynogenesis or paternal, androgenesis) will not develop to birth. The explanation for this observation is that some genes are imprinted, either paternally or maternally. Hence, in gynogenetic individuals, maternally imprinted genes do not have an active copy. Similarly, androgenetic individuals do not have an active copy for paternally imprinted genes. Lacking functionally active copy of these imprinted genes results in lethal defects and hence those individuals cannot survive to term. Well regulated imprinted genes are necessary for normal development and dysregulated imprinting mechanisms are believed to be the cause of some genetic conditions (Hall, 1990; Solter, 1992; Falls *et al.*, 1999; Clayton-Smith, 2003; Wilkins and Ubeda, 2011; Horsthemke, 2014; Peters, 2014). In humans, or example, Prader-Willi (PWS) and Angelman (AS) syndromes are sister imprinting-induced disorders involving deletion of DNA segments derived from different parents at the same genomic region (Meijers-Heijboer *et al.*, 1992; Nicholls *et al.*, 1998; Cassidy *et al.*, 2000). In livestock and crops, on the other hand, individuals with defects caused by malfunctioning imprinted genes are less likely to be observed since they tend to be culled by natural selection. However, imprinted genes often contribute to many economically important traits in agriculture (Wolf *et al.*, 2008; Spencer, 2009; Lawson *et al.*, 2013; Gehring, 2013). In sheep, for instance, the Callypyge locus can result in a muscular hindquarter phenotype only when the *CLPG* mutation was inherited from the sire (Georges *et al.*, 2003). In cattle, well regulated imprinted genes *CDKN1C* and *PHLDA2* are critical for the developmental status and the quality of the embryo (Driver *et al.*, 2013). In maize, corn kernels are fully colored when alleles of R and B genes are inherited from the mother but variegated if the alleles are paternally inherited (Kermicle, 1970; Selinger and Chandler, 2001), which gave the first description of differential contribution of parental genomes in plants. In *Arabidopsis*, individuals inheriting excess paternal or maternal

genomes display reciprocal size phenotypes (Scott *et al.*, 1998). Cumulative evidence has suggested that imprinting has an essential role on endosperm development of flowering plants (Vinkenoog *et al.*, 2003). Therefore, it may have a big impact on yield traits of many crop species, such as corn and rice.

2.2.2 Known Mechanisms of Genomic Imprinting

The prevalence and potential importance of imprinted genes have stimulated a constellation of intensive studies on the mechanisms of imprinting. Although the precise mechanism has not been fully understood yet, it is widely believed that differential DNA methylation on parental genomes is the main cause of imprinting in both plants and animals (Razin and Cedar, 1994; McEwen and Ferguson-Smith, 2009).

There are three main stages in the proposed methylation model of genomic imprinting: 1) the acquisition and maintenance of imprinting established from the previous generation; 2) the erasure and resetting of imprints in the germ cells of the current generation; 3) the establishment of imprinting for the next generation (Figure 2.5). In the first stage, when the embryo is diploid after fertilization, the imprints are maintained in cells of the embryo, yolk sac, placenta, and also in the adult after each cell division on the same parental chromosome as acquired from the gametes. In the second stage, germ cells are formed in the embryonic gonad and the imprints from the previous generation are erased by a genome-wide demethylation completed by embryonic day 12$\sim$13 in both sexes (Brandeis *et al.*, 1993; Tada *et al.*, 1998). Then in the third stage, during the development of germ line into sperm or egg cells after sex determination, new imprints are established such that in mature gametes they reflect the sex of that germ line for the next generation. This stage is triggered by *de novo* methylation that starts in both germ lines at fetal stage and continues after birth (Kafri *et al.*, 1992; Brandeis *et al.*, 1993).

Once established, genomic imprinting in somatic cells is regulated by imprinting control regions (ICR) in *cis*. ICRs are often associated with differentially methylated regions (DMR) that reside in

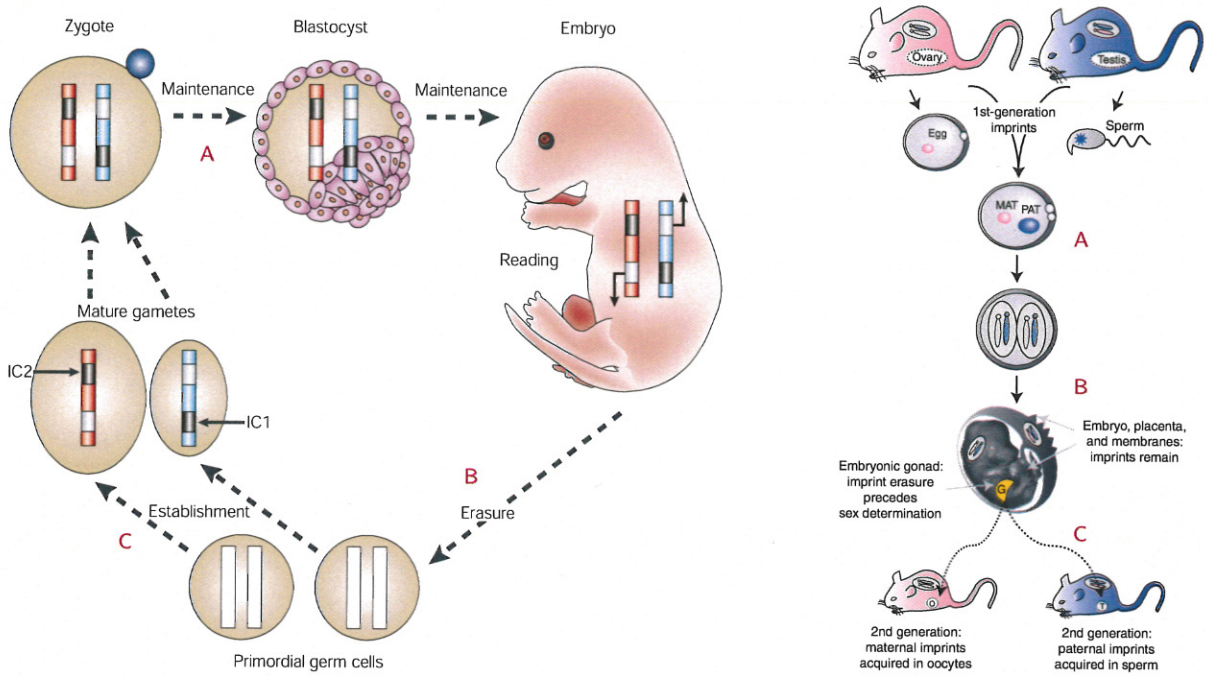


Figure 2.5: Life cycle of methylation imprints in mammalian development (a: Reik and Walter, 2001; b: Barlow and Bartolomei, 2014). The three stages of imprinting in the methylation model are A: acquisition and maintenance of imprinting from previous generation; B: erasure and resetting of imprints in the germ cells of the current generation; and C: establishment of imprinting for the next generation.

different locations of the genome (e.g., gene promoter region). DMRs are differentially methylated on DNA from different parents based on sex recognition mechanisms, although largely remain unclear at the current stage, during gametogenesis in the parental generation described above. Because DNA methylation is highly associated with gene transcription, expression repression as a result of the *cis*-regulation of ICRs becomes observable genomic imprinting in somatic cells. Since DNA methylation is featured by its reversibility, erasure of imprints from previous generation in the embryonic gonad of the current generation is possible and hence this mechanism cycles from generation to generation.

Besides this widely accepted methylation model on imprinting mechanism, other hypotheses include the insulator model and the non-coding RNA model (Ideraabdullah *et al.*, 2008). The insulator model was established based on the observation of the paternally imprinted *H19* gene

and the maternally imprinted *Igf2* gene. These two genes, physically close to each other, reside on chromosome 11 in humans and are found in conserved synteny on distal chromosome 7 in mice (Bartolomei *et al.*, 1991; DeChiara *et al.*, 1991). These two genes are regulated by a shared ICR, which locates approximately 2 kb upstream of *H19* and acts by regulating the interactions between the *H19* and *Igf2* promoters and their shared enhancers, which lie downstream of *H19* (Kaffer *et al.*, 2000, 2001; Phillips and Corces, 2009). The proper imprinting of *H19* and *Igf2* requires this ICR to be methylated on the paternal allele and unmethylated on the maternal allele (Jinno *et al.*, 1996; Tremblay *et al.*, 1997). One characteristic of this ICR is that this sequence will bind to the insulator protein CCCTC-binding factor (CTCF) when it is unmethylated (Bell and Felsenfeld, 2000; Hark *et al.*, 2000; Kaffer *et al.*, 2000; Kanduri *et al.*, 2000; Szabó *et al.*, 2000). Therefore, binding of CTCF to the unmethylated maternal ICR protects it from *de novo* methylation and prevents downstream enhancers from activating *Igf2*, but leaving them available to activate *H19*. On the other hand, CTCF is unable to bind the methylated paternal ICR and hence resulting in the expression of *Igf2* while *H19* is silenced (Szabó *et al.*, 2004, Figure 2.6).

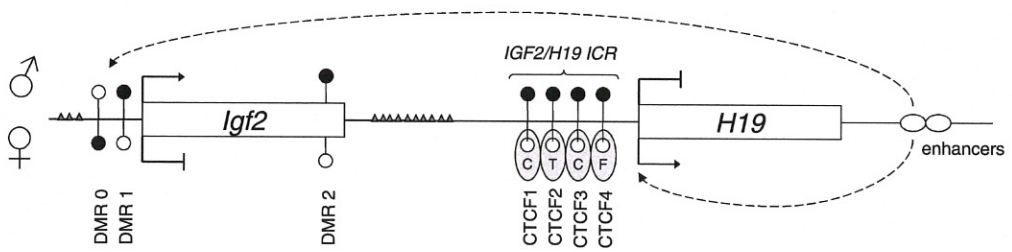


Figure 2.6: Regulation of the *H19/Igf2* locus (Pidsley *et al.*, 2012). The shared ICR is located between *H19* and *Igf2* and contains four CTCF binding sites (CTCF1~4). This ICR is normally methylated (filled circles) on the paternal allele (top) and unmethylated (empty circles) on the maternal allele (bottom). The enhancers will then interact with gene promoters (dashed arrows pointing from right to left) according to the binding status of CTCF and determine the active status of each gene (solid arrows at genes pointing from left to right).

The function of ncRNA in regulating imprinting can be demonstrated from the observation on the *Kcnq1* locus, which is immediately adjacent to the *H19/Igf2* locus in both mouse and human. This locus harbors a paternally expressed gene *Kcnq1ot1* that encodes a lncRNA and several maternally expressed protein-coding genes, including *Cdkn1c*, *Ascl2*, *Phlda2*, and others (Thorvaldsen

and Bartolomei, 2007). The promoter of the *Kcnq1ot1* gene resides within KvDMR1, the ICR governing the *Kcnq1* locus. Hypermethylation on the maternal allele represses the lncRNA and activates the adjacent protein-coding imprinted genes, whereas hypomethylation of KvDMR1 on the paternal allele is associated with *Kcnq1ot1* expression and repression of the adjacent imprinted genes, which may result in the Beckwith-Wiedemann Syndrome (Diaz-Meyer *et al.*, 2003). Deletion of KvDMR1 on the paternal allele results in a failure to express *Kcnq1ot1* and in biallelic expression of the genes that are normally expressed on the maternal allele (Fitzpatrick *et al.*, 2002; Mancini-Dinardo *et al.*, 2006), suggesting that transcription of the ncRNA Kcnq1ot1 is essential to silence the adjacent protein-coding genes in *cis*.

It is also worth noting that, although XCI randomly silences one of the two X-chromosomes for dosage compensation of the X-linked genes in males and females, some studies suggested that imprinting may exist in some region of the X-chromosome (Monk and Grant, 1990; Cattanach and Beechey, 1990; Iwasa and Pomiankowski, 1999). Random XCI is often found in the postimplantation embryo, whereas imprinted XCI of the paternal X-chromosome (Xp) has been observed in the preimplantation embryo (Huynh and Lee, 2003; Mak *et al.*, 2004; Okamoto *et al.*, 2004) and the extraembryonic lineages (Takagi and Sasaki, 1975; West *et al.*, 1977, 1978) in mouse, in the placenta of bovine (Xue *et al.*, 2002), and in both extraembryonic and somatic tissues in marsupials (Cooper *et al.*, 1971, 1983; Samollow *et al.*, 1995). The mechanism of imprinted XCI has not been fully understood yet, but evidence suggested that it may not rely on DNA methylation in the same manner as autosomal imprinted genes (Ideraabdullah *et al.*, 2008), and imprinted XCI may have evolved independently in eutherians and marsupials, or eutherians may have developed new mechanisms to initiate XCI in the early embryo from a template mechanism still used in marsupials (Huynh and Lee, 2005). Since there is so far no strong evidence showing Xp is preferentially inactivated in extraembryonic tissues in human (Vasques *et al.*, 2002), plus that human XmXmXp females and XmXmY males have less severe developmental defects than their mouse counterparts (MacDonald *et al.*, 1994), it is suggested that the existence of imprinted XCI is unlikely in humans

and, in contrast to autosomal imprinted genes where imprinting is widely conserved, imprinted XCI appears species-specific (Ideraabdullah *et al.*, 2008).

In summary, with evidence observed so far, DNA methylation is the main cause of most genomic imprinting instances. However, other epigenetic mechanism are also responsible for some imprinted genes. Since more imprinted genes may be discovered in the future, other unrevealed paradigms for imprinting regulation may exist and hypotheses of mechanisms of imprinting may get more complicated to provide a better understanding of this epigenetic process.

2.2.3 Key Features of Genomic Imprinting

Although the mechanisms of genomic imprinting are still under intensive investigations, observations obtained so far provide a list of characteristics of imprinted genes. Apart from the fact that imprinting is observed mainly in mammals and angiosperms, which lead to the famous "parent-offspring conflict" hypothesis of the emergence of imprinting, the following features may give a closer link between the observations and the molecular mechanism underlying imprinting.

Non-random Monoallelic Expression

The most important feature that distinguishes genomic imprinting from other epigenetic phenomena (especially gene silencing process) is its non-randomness. As introduced before, both XCI and paramutation are epigenetic-related gene inactivation observations. However, neither is featured by a preferential inactivation of a specific allele in a parent-of-origin manner, except few observations on imprinted XCI. Recently, widespread monoallelic expression was observed on human autosomes (Gimelbrant *et al.*, 2007). About 20% of genes in the genome are subject to this expression pattern, but interestingly, this monoallelic expression is random and is defined as random monoallelic expression (RMAE). As more evidence on RMAE accumulates (Zwemer *et al.*, 2012; Tang *et al.*, 2011; Eckersley-Maslin *et al.*, 2014; Deng *et al.*, 2014; Gendrel *et al.*, 2014; Li *et al.*, 2012), the current

understanding of gene expression regulation may experience a huge change but unfortunately, little is known about the links between RMAE and genomic imprinting (Gregg, 2014).

Clustering

Current census of imprinted genes showed that around 80% are physically linked in clusters with other imprinted genes (Reik and Walter, 2001; Verona *et al.*, 2003; Wan and Bartolomei, 2008). These clusters contain two or more imprinted genes over a region that can span 1 megabase (Mb) or more (Thorvaldsen and Bartolomei, 2007). Taking the mouse (*Mus musculus*) as an example, many imprinted genes are found on chromosome 7, on which there are 5 big clusters and various number of imprinted genes are located in each clusters (Figures 2.7 and 2.8). The genes in clusters, which can be either maternally or paternally imprinted, are usually jointly regulated through a common ICR that can act over distances of a megabase or more (Yang *et al.*, 1998).

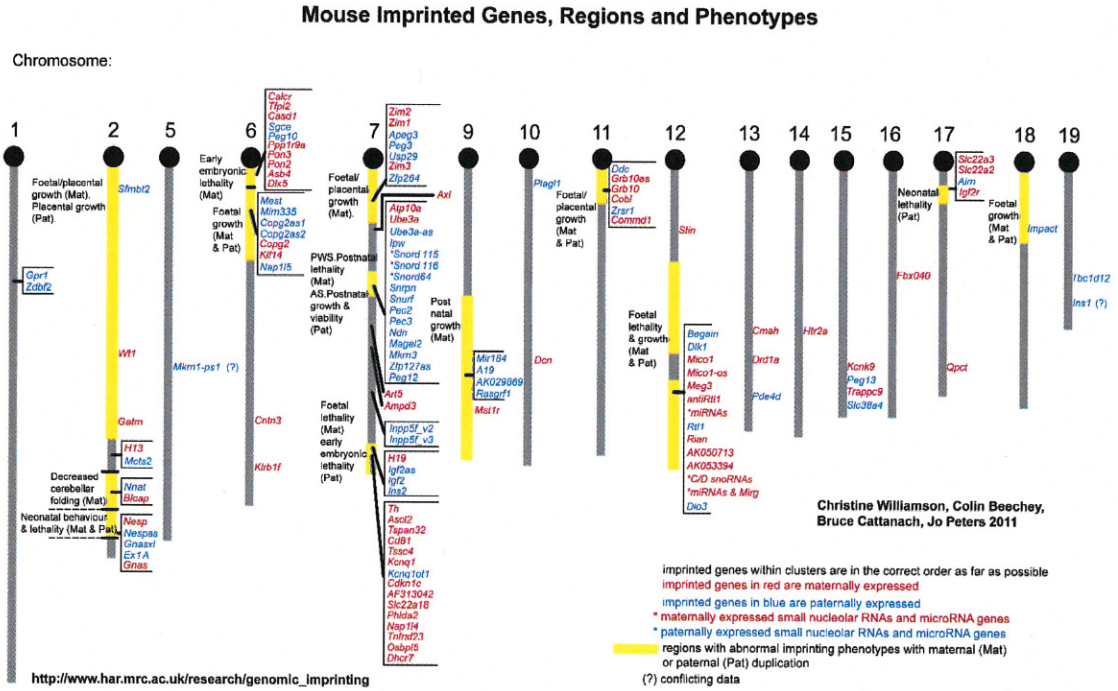


Figure 2.7: Imprinted gene clusters and their locations in the mouse genome (Williamson *et al.*, 2013; `http://www.mousebook.org`, last updated in 2011).

The aforementioned *H19/Igf2* locus in both humans and mice is probably the most well-studied

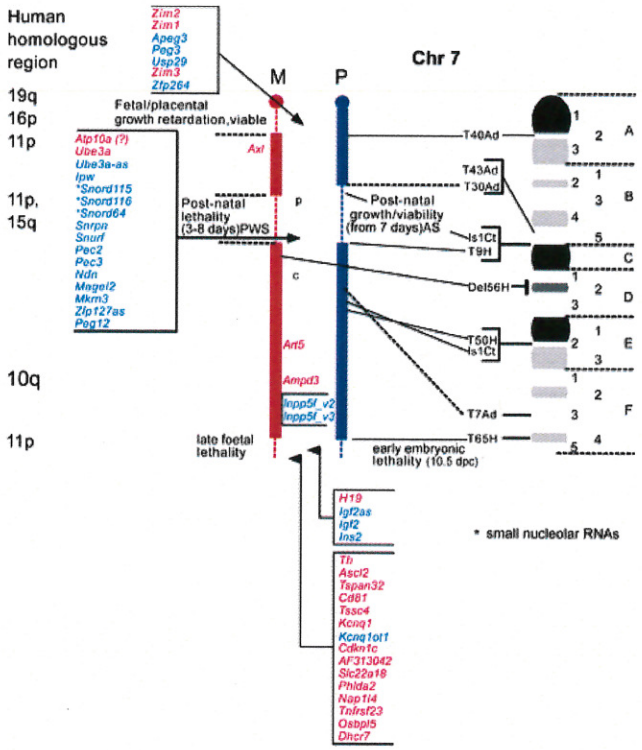


Figure 2.8: Imprinted gene clusters on murine chromosome 7 along with human homologous regions on chromosome 11 (Williamson *et al.*, 2013; `http://www.mousebook.org`).

imprinted cluster. It locates at 11p15.5 in humans and 7pF5 in mice and the two genes are approximately 163 and 72 kb apart from each other in the two species, respectively. The shared ICR, usually designated as imprinting center 1 (IC1) in humans and differentially methylated domain (DMD) in mice, locates between the two genes and is approximately 5 kb and 2 kb long, respectively (Jinno *et al.*, 1996; Tremblay *et al.*, 1997). Well regulated ICR in this region is required for normal development while a hypomethylation of this ICR can result in the epigenetic silencing of both *IGF2* alleles in humans (Netchine *et al.*, 2007; Abu-Amero *et al.*, 2010). This will lead to the Silver-Russell syndrome (SRS), a gestational developmental disorder defined by intrauterine growth restriction (IUGR) that shows lowered birth weight, dwarfness at adulthood, and other abnormalities like feeding problems, lack of subcutaneous fat, early onset of puberty, and so forth (Silver *et al.*, 1953; Russell, 1954; Wollmann *et al.*, 1995; Abu-Amero *et al.*, 2008; Penaherrera *et al.*, 2010). Another well-studied imprinting cluster is the *Kcnq1* locus introduced before to

elucidate the function of ncRNA in imprinting mechanism. This locus is physically adjacent to the *H19/Igf2* cluster at 11p15.4 in humans and 7qF5 in mice. KvDMR1, the shared ICR, is in charge of regulating one paternally expressed ncRNA and nine maternally expressed genes in this cluster (Thorvaldsen and Bartolomei, 2007).

It is not clear why imprinted genes tend to be in clusters. However, some common features are shared by many imprinted clusters characterized so far. For example, many imprinted clusters contain several imprinted protein-coding genes and at least one imprinted ncRNA (Thorvaldsen and Bartolomei, 2007). Interestingly, it was found that the imprinting directions of the imprinted genes in one cluster are not always the same. Instead, it was often observed that one parental chromosome expresses the ncRNA and the other expresses the various mRNA genes (Barlow, 2011). This may indicate that an imprinted ncRNA and mRNAs cannot normally be expressed in *cis*, and that the expression of the ncRNA might be part of the silencing mechanism in imprinted gene clusters (Barlow, 2005).

Tissue- and Developmental-Stage-Specificity

Epigenetic gene regulation plays an important role in cell and tissue differentiation. Because imprinting is a specialized form of gene regulation, some features of gene regulation are conserved in genomic imprinting, among which tissue-specific and developmental-stage-specific patterns are the two most commonly noticed ones.

In a recent review (Prickett and Oakey, 2012), 82 imprinted genes published in the Web Atlas of Murine genomic Imprinting and Differential EXpression (WAMIDEX, Schulz *et al.*, 2008) were surveyed. The expression of these genes has been tested in multiple tissues and a large proportion (23 genes, 28%) of these genes showed tissue-specific imprinting status in only one specific tissue type. Within these 23 genes, the majority are imprinted only in extra-embryonic tissues, specifically placenta (48%), and yolk sac (9%), or only in the brain, including specific subsets of brain regions

(39%). The only published example of a tissue-specific imprinted gene where imprinting is not localized to the brain or extra-embryonic tissue is Dopa decarboxylase (*Ddc*), which is imprinted solely in heart (Menheniott *et al.*, 2008).

Besides this subset of imprinted genes, another subset is almost ubiquitously imprinted in many tissues and yet bi-allelically expressed in a single tissue. This was first observed at the *Igf2* locus that it reverts to bi-allelic expression in the choroid plexus and leptomeninges (DeChiara *et al.*, 1991). The biological function of this reversion is largely unknown, but recent studies on the *Dlk1* gene suggestes that it may play an important role in the regulation of gene dosage (Ferron *et al.*, 2011). Also, it was interesting to observe that during development, the growth factor receptor-binding protein gene *Grb10* is paternally expressed in the brain with a function of tempering social dominance (Garfield *et al.*, 2011) while maternally expressed in most other peripheral tissues to repress growth and in adulthood mediates glucose metabolism and energy homeostasis (Smith *et al.*, 2007; Charalambous *et al.*, 2003, 2010).

Differential methylation between parental alleles is known to control imprinted gene expression (Reik and Walter, 2001). Therefore, genome-wide identification of variations in different tissue types and in DMRs, particularly somatic DMRs, could be potentially indicative for tissue-specific imprinting. For this purpose, the G_s protein alpha-subunit gene ($G_s\alpha$) is mentioned here as an example. $G_s\alpha$ locates in the *GNAS* cluster that consists of multiple different imprinted transcripts (Peters and Williamson, 2007) and is maternally expressed in a specific subset of tissues composed of renal cortex, brown and white adipose tissue (Yu *et al.*, 1998), the anterior pituitary (Hayward *et al.*, 2001), thyroid (Mantovani *et al.*, 2002; Germain-Lee *et al.*, 2002), ovary (Mantovani *et al.*, 2002), and paraventricular nucleus of the hypothalamus (Chen *et al.*, 2009), but bi-allelically expressed in most other tissues. This "dual form" of being either monoallelic or biallelic expression pattern motivated intensive studies on the ICR of the $G_s\alpha$ gene. The *1A* DMR is known to control imprinting in these tissues (Williamson *et al.*, 2004). However, no tissue-specific differences in methylation at this DMR have so far been found, and thus tissue-specific imprinting may involve

more than simple tissue-specific methylation differences and it might be important to consider other epigenetic variations (Prickett and Oakey, 2012).

As another characteristic, imprinted genes show a high developmental-stage-specificity (Wood and Oakey, 2006). For example, the mouse *Murr1* gene has been reported to be biallelically expressed in the whole body of neonatal mice specimens (Nabetani *et al.*, 1997), but was verified to be maternally expressed in adult tissues, especially in brain (Wang *et al.*, 2004). Also, the *IPL* gene (*i*mprinted in *p*lacenta and *l*iver) was reported to show developmental-stage-specific imprinted expression in both mice (Qian *et al.*, 1997) and domestic pigs (Hou *et al.*, 2010). Besides autosomal imprinted genes, developmental-stage-specificity is also shared by some X-linked imprinted genes, for instance, the *Xlr* gene (Raefski and O'Neill, 2005). However, the evolutionary implications of this developmental-stage-specificity remains largely unclear (Gregg, 2014) and more investigations are needed for a better understanding of this feature.

2.2.4 The Emergence and Evolution of Genomic Imprinting

Although the picture of molecular mechanisms of genomic imprinting has become clearer with recent studies, the evolutionary survival of imprinting under the pressure of natural selection has always been a puzzle to geneticists. The main reason is that, the functional haploidy caused by imprinting may increase the risk of being exposed to a deleterious mutation, unlike the case of Mendelian inheritance where there is always a "backup" compensation from the other allele. Therefore, the fitness gain from imprinting must overwhelm this fitness cost to keep imprinting not eliminated by natural selection. This topic has attracted a broad interest, and in a recent issue of the journal Heredity (August 2014, Volume 113, issue 2), all articles are on evolutionary theories of imprinting. Up to date, more than 15 hypotheses have been proposed in an attempt to solve this mystery. Most of these hypotheses applied only to certain specific cases and very few have been generally accepted. Thus, only the most influential hypotheses are highlighted here.

Parent-offspring Conflict Hypothesis

Parent-offspring conflict (or more generally kinship) hypothesis (Haig and Westoby, 1989; Moore and Haig, 1991; Haig, 2000) is probably the most widely accepted and empirically supported hypothesis presented so far. Given the fact that imprinting was mainly observed in mammals and flowering plants where there is always a nutrition connection between mother and offspring, it was suggested that imprinting is a mechanism regarding the competition between paternally expressed genes and maternally expressed genes on the allocation of maternal resources contributed to the development of offspring.

Take mammals as an example, the mother and the offspring are connected through the placenta and the umbilical cord. All nutrition demands required for the development of the fetus come from this connection. Assuming that the females can have offspring from different males in the same pregnancy (multi-paternity), alleles inherited from the father tend to extract as much nutrition as possible from the mother to promote the development of the fetus such that his own offspring could be more competitive after birth. The females, on the other hand, will treat all her offspring equally because they all equally related to herself, even if the offspring may come from different fathers. Thus, maternal investments are spread to different fetuses as evenly as possible by maternally derived genes. Moreover, maternally inherited genes will tend to limit the nutritional investment because resources need to be conserved for her future pregnancy, resulting in that nutrition offered to offspring will be restricted at a sufficient-for-development level. Therefore, there is a conflict between paternally and maternally derived genes on the allocation of maternal resources to offspring, from which the hypothesis was named. The case in flowering plants is very similar to that observed in mammals where seeds (offspring) and mother are directly connected. Hence, it was predicted that genes that increase maternal provisioning are paternally expressed (maternally imprinted) whereas the opposite is predicted for genes that decrease maternal provisioning.

The parent-offspring conflict hypothesis successfully explained what had been observed up to

that moment and was well supported by experimental results. For example, the growth enhancers *Igf2* and *Peg3* are both paternally expressed and are conservative in many eutherian species (Haig and Graham, 1991; Li *et al.*, 1999; Smits *et al.*, 2008), whereas the growth inhibitors are mostly maternally expressed (Haig, 1997; Burt and Trivers, 1998; Spencer *et al.*, 1998). This hypothesis also suggested that the direct connection between mother and offspring might be necessary for imprinting, which would explain why imprinting was not observed in oviparous species, for example, birds (Frésard *et al.*, 2013). The explanation was that for these species, the amount of resources for the development of the embryo has been determined after, say, the laying of the fertilized egg, which is everything included inside the shell. Thus, absorbing resources from the mother for paternally derived gene and retaining resources for maternally derived gene is not needed.

Maternal-Offspring Coadaptation Hypothesis

Another hypothesis on the evolution of imprinting is called the maternal-offspring coadaptation hypothesis (Wolf and Hager, 2006), where instead of conflict, this hypothesis suggested that the expression of maternally derived alleles allows for the coadaptation of complementary traits between mother and offspring and enhances offspring's development and fitness. This hypothesis assumes that both maternal and offspring phenotypes affect offspring fitness interactively, and hence predicts that more maternally expressed genes should be observed than paternally expressed genes, which is the case for genes that are exclusively imprinted in the placenta where all genes are expressed from the maternal allele (Coan *et al.*, 2005; Wagschal and Feil, 2006). Also, the coadaptation hypothesis predicts that the incidence of imprinting should be higher in taxa where mother-offspring interactions have a greater effect on offspring fitness.

This is supported by the evidence that mouse pups are better provisioned by foster mothers of the same strain than their natural mothers, suggesting a coacaptation between offspring and maternal phenotype (Hager and Johnstone, 2003, 2005). Given their assumptions and the empirical observations, Wolf and Hager (2006) mathematically analyzed the relationship between the level

of imprinting and the average fitness of individuals. They found that imprinting will be favored when genetic variation exists for coadapted traits, and concluded that this genetic variation will influence the evolution of genomic imprinting because the mean fitness of the population will be increased by coadaptation.

Intralocus Sexual Conflict Hypothesis

Both the parent-offspring conflict hypothesis and the coadaptation hypothesis assumed that each individual in a population is under the same selection pressure. However, if selection is sexually antagonistic for a gene, either in magnitude or in direction, alleles enhancing female benefit will be favored in matrigenes and alleles enhancing male benefit will be favored in patrigenes. Therefore, natural selection would favor the so-called "modifier loci" that silence maternal alleles in males and that silence paternal alleles in females, resulting in that imprinted genes being selected in a sex-specific way called sexually dimorphic imprinting (Bonduriansky, 2007). This hypothesis is known as the intralocus sexual conflict hypothesis on the evolution of genomic imprinting (Day and Bonduriansky, 2004). As a result, in sex-specific imprinting, the expression of an allele depends not only on the sex of the parent, but also on the sex of the recipient offspring.

Day and Bonduriansky (2004) validated their hypothesis by a mathematical model deduction and supports for this hypothesis have been reported (Gregg *et al.*, 2010a). Also, this hypothesis predicts that many sexually selected traits should show imprinting, which is consistent with the fact that both growth and behavior can be sexually selected (Ashbrook and Hager, 2013). However, other studies found that sex-dependent imprinting, although may exist, does not unequivocally support the prediction of this sexual antagonism hypothesis (Hager *et al.*, 2008), and more convincing evidence needs to be provided in the future (Moore and Mills, 2008; Spencer and Clark, 2014). Despite this, the sexual antagonism hypothesis predicts some evolutionary machinery of imprinting given a potential differential selection pressure applied on different sexes (Patten *et al.*, 2014).

2.2.5 Many Unknowns of Genomic Imprinting

Since the first discovery of genomic imprinting thirty years ago, scientific efforts have been made for a better understanding of this epigenetic regulation mechanism. Although significant progress has been reached, many aspects of imprinting still remain unknown (Gregg, 2014). For example, the aforementioned random monoallelic inactivation observed in human autosomes might be related to imprinting, but the underlying mechanism and its functions are largely uncovered. Also, the "full-null" definition of genomic imprinting based on monoallelic expression has been challenged since recent studies have provided evidence that for some imprinted loci, both alleles are yet differentially expressed in a parent-of-origin-preferential or parent-of-origin-dependent manner (Khatib, 2007), indicating that silencing is incomplete (Abramowitz and Bartolomei, 2009; Morcos *et al.*, 2011; Barlow, 2011). This divergence from canonical monoallelic expression imprinting mechanism, together with imprinted XCI, may change our definition of genomic imprinting, and may also change our predictions on the total number of imprinted genes in the whole genome. Numerous hypotheses on the evolution of imprinting have been proposed, but still, no one offers a general solution to the question of why imprinting is favored by natural selection. The origin of imprinting has been examined by means of molecular evolution (McVean and Hurst, 1997; Smith and Hurst, 1998; Parker-Katiraee *et al.*, 2007; Spillane *et al.*, 2007; Miyake *et al.*, 2009; Hutter *et al.*, 2010; O'Connell *et al.*, 2010; Wolff *et al.*, 2011), but no consistent agreement has been reached.

In summary, the discovery of genomic imprinting has changed views of Mendelian inheritance. The different mechanisms of imprinting uncovered so far have been mainly based on observed imprinting incidences, and the proposed evolutionary hypotheses on imprinting have relied largely on empirical observations as well. With more intensive studies and more advanced techniques in the biological sciences, it can be expected that a clearer picture of the function, mechanism, and evolution of imprinting will emerge.

References

Abramowitz, L. K. and M. S. Bartolomei, 2009. An in vitro ES cell imprinting model shows that imprinted expression of the *Igf2r* gene arises from an allele-specific expression bias. Development, 136: 437–448

Abu-Amero, S., D. Monk, J. Frost, *et al.*, 2008. The genetic aetiology of Silver-Russell syndrome. J. Med. Genet., 45(4): 193–199

Abu-Amero, S., E. L. Wakeling, M. Preece, *et al.*, 2010. Epigenetic signatures of Silver-Russell syndrome. J. Med. Genet., 47(3): 150–154

Alleman, M., L. Sidorenko, K. McGinnis, *et al.*, 2006. An RNA-dependent RNA polymerase is required for paramutation in maize. Nature, 442(7100): 295–298

Anaka, M., A. Lynn, P. McGinn, *et al.*, 2009. Genomic imprinting in *Drosophila* has properties of both mammalian and insect imprinting. Dev. Genes Evol., 219(2): 59–66

Ashbrook, D. G. and R. Hager, 2013. Empirical testing of hypotheses about the evolution of genomic imprinting in mammals. Front. Neuroanat., 7: 6

Barlow, D., 2005. Mechanisms of monoallelic gene expression. Horizon Symposia 6 (Beyond the genome: epigenetic regulation of identity), http://www.nature.com/horizon/epigenetics/kq/2_Barlow.html

Barlow, D. P., 2011. Genomic Imprinting: A Mammalian Epigenetic Discovery Model. Annu. Rev. Genet., 45: 379–403

Barlow, D. P. and M. S. Bartolomei, 2014. Genomic imprinting in mammals. Cold Spring Harb. Perspect. Biol., 6(2): a018382

Barlow, D. P., R. Stoger, B. G. Herrmann, *et al.*, 1991. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the *Tme* locus. Nature, 349(6304): 84–87

Bartolomei, M. S., S. Zemel, and S. M. Tilghman, 1991. Parental imprinting of the mouse *H19* gene. Nature, 351(6322): 153–155

Barton, S. C., M. A. Surani, and M. L. Norris, 1984. Role of paternal and maternal genomes in mouse development. Nature, 311(5984): 374–376

Bean, C. J., C. E. Schaner, and W. G. Kelly, 2004. Meiotic pairing and imprinted X chromatin assembly in Caenorhabditis elegans. Nat. Genet., 36(1): 100–105

Bell, A. C. and G. Felsenfeld, 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. Nature, 405(6785): 482–485

Bell, C. G., 2013. Epigenome-Wide Association Studies: Potential Insights into Human Disease. In A. Naumova and C. Greenwood (Editors), *Epigenetic and Complex Traits*, pp. 287–317. Springer

New York

Berger, S. L., 2002. Histone modifications in transcriptional regulation. Curr. Opin. Genet. Dev., 12(2): 142–148

Bestor, T. H. and G. L. Verdine, 1994. DNA methyltransferases. Curr. Opin. Cell Biol., 6(3): 380–389

Bird, A., 2007. Perceptions of epigenetics. Nature, 447(7143): 396–398

Bird, A. P., 1984. DNA methylation versus gene expression. J. Embryol. Exp. Morphol., 83 (Suppl.): 31–40

Bird, A. P. and A. P. Wolffe, 1999. Methylation-induced repression–belts, braces, and chromatin. Cell, 99(5): 451–454

Bock, C. and T. Lengauer, 2008. Computational epigenetics. Bioinformatics, 24(1): 1–10

Bonduriansky, R., 2007. The genetic architecture of sexual dimorphism: the potential roles of genomic imprinting and condition-dependence. In J. F. Daphne, U. B. Wolf, and T. Székely (Editors), *Sex, Size and Gender Roles: Evolutionary Studies of Sexual Size Dimorphism*, pp. 176–184. Oxford University Press, Oxford, UK

Brandeis, M., T. Kafri, M. Ariel, *et al.*, 1993. The ontogeny of allele-specific methylation associated with imprinted genes in the mouse. EMBO J., 12(9): 3669–3677

Brink, R. A., E. D. Styles, and J. D. Axtell, 1968. Paramutation: directed genetic change. Paramutation occurs in somatic cells and heritably alters the functional state of a locus. Science, 159(3811): 161–170

Burt, A. and R. Trivers, 1998. Genetic conflicts in genomic imprinting. Proc. Biol. Sci., 265(1413): 2393–2397

Cassidy, S. B., E. Dykens, and C. A. Williams, 2000. Prader-Willi and Angelman syndromes: sister imprinted disorders. Am. J. Med. Genet., 97(2): 136–146

Cattanach, B. M. and C. V. Beechey, 1990. Autosomal and X-chromosome imprinting. Development (Suppl.), pp. 63–72

Charalambous, M., M. Cowley, F. Geoghegan, *et al.*, 2010. Maternally-inherited *Grb10* reduces placental size and efficiency. Dev. Biol., 337(1): 1–8

Charalambous, M., F. M. Smith, W. R. Bennett, *et al.*, 2003. Disruption of the imprinted *Grb10* gene leads to disproportionate overgrowth by an *Igf2*-independent mechanism. Proc. Natl. Acad. Sci. U.S.A., 100(14): 8292–8297

Chen, M., J. Wang, K. E. Dickerson, *et al.*, 2009. Central nervous system imprinting of the g protein $G_s\alpha$ and its role in metabolic regulation. Cell Metabolism, 9(6): 548–555

Chen, Z. X. and A. D. Riggs, 2011. DNA methylation and demethylation in mammals. J. Biol.

Chem., 286(21): 18347–18353

Cheng, X., 1995. Structure and function of DNA methyltransferases. Annu Rev Biophys Biomol Struct, 24: 293–318

Cheung, P. and P. Lau, 2005. Epigenetic regulation by histone methylation and histone variants. Mol. Endocrinol., 19(3): 563–573

Clayton-Smith, J., 2003. Genomic imprinting as a cause of disease. BMJ, 327(7424): 1121–1122

Coan, P. M., G. J. Burton, and A. C. Ferguson-Smith, 2005. Imprinted genes in the placenta–a review. Placenta, 26 (Suppl. A): 10–20

Cooper, D. W., J. L. VandeBerg, G. B. Sharman, *et al.*, 1971. Phosphoglycerate kinase polymorphism in kangaroos provides further evidence for paternal X inactivation. Nature New Biol., 230(13): 155–157

Cooper, D. W., P. A. Woolley, G. M. Maynes, *et al.*, 1983. Studies on metatherian sex chromosomes. XII. Sex-linked inheritance and probable paternal X-inactivation of alpha-galactosidase A in Australian marsupials. Aust. J. Biol. Sci., 36(5-6): 511–517

Day, T. and R. Bonduriansky, 2004. Intralocus sexual conflict can drive the evolution of genomic imprinting. Genetics, 167(4): 1537–1546

DeChiara, T. M., E. J. Robertson, and A. Efstratiadis, 1991. Parental imprinting of the mouse insulin-like growth factor II gene. Cell, 64(4): 849–859

Deng, Q., D. Ramskold, B. Reinius, *et al.*, 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science, 343(6167): 193–196

Diaz-Meyer, N., C. D. Day, K. Khatod, *et al.*, 2003. Silencing of *CDKN1C* ($p57^{KIP2}$) is associated with hypomethylation at *KvDMR1* in Beckwith-Wiedemann syndrome. J. Med. Genet., 40(11): 797–801

Driver, A. M., W. Huang, J. Kropp, *et al.*, 2013. Knockdown of *CDKN1C* ($p57^{kip2}$) and *PHLDA2* results in developmental changes in bovine pre-implantation embryos. PLoS ONE, 8(7): e69490

Eckersley-Maslin, M. A., D. Thybert, J. H. Bergmann, *et al.*, 2014. Random monoallelic gene expression increases upon embryonic stem cell differentiation. Dev. Cell, 28(4): 351–365

Ehrlich, M., G. G. Wilson, K. C. Kuo, *et al.*, 1987. N4-methylcytosine as a minor base in bacterial DNA. J. Bacteriol., 169(3): 939–943

Esteller, M., 2008. Epigenetics in Cancer. N. Engl. J. Med., 358(11): 1148–1159

Falls, J. G., D. J. Pulford, A. A. Wylie, *et al.*, 1999. Genomic imprinting: implications for human disease. Am. J. Pathol., 154(3): 635–647

Feil, R. and F. Berger, 2007. Convergent evolution of genomic imprinting in plants and mammals. Trends Genet., 23(4): 192–199

Ferron, S. R., M. Charalambous, E. Radford, *et al.*, 2011. Postnatal loss of Dlk1 imprinting in stem cells and niche astrocytes regulates neurogenesis. Nature, 475(7356): 381–385

Fitzpatrick, G. V., P. D. Soloway, and M. J. Higgins, 2002. Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. Nat. Genet., 32(3): 426–431

Frésard, L., M. Morisson, J. M. Brun, *et al.*, 2013. Epigenetics and phenotypic variability: some interesting insights from birds. Genet. Sel. Evol., 45: 16

Gardiner-Garden, M. and M. Frommer, 1987. CpG islands in vertebrate genomes. J. Mol. Biol., 196(2): 261–282

Garfield, A. S., M. Cowley, F. M. Smith, *et al.*, 2011. Distinct physiological and behavioural functions for parental alleles of imprinted *Grb10*. Nature, 469(7331): 534–538

Gehring, M., 2013. Genomic imprinting: insights from plants. Annu. Rev. Genet., 47: 187–208

Gendrel, A. V., M. Attia, C. J. Chen, *et al.*, 2014. Developmental dynamics and disease potential of random monoallelic gene expression. Dev. Cell, 28(4): 366–380

Georges, M., C. Charlier, and N. Cockett, 2003. The callipyge locus: evidence for the *trans* interaction of reciprocally imprinted genes. Trends Genet., 19(5): 248–252

Germain-Lee, E. L., C. L. Ding, Z. Deng, *et al.*, 2002. Paternal imprinting of $G\alpha_s$ in the human thyroid as the basis of TSH resistance in pseudohypoparathyroidism type 1a. Biochem. Biophys. Res. Commun., 296(1): 67–72

Gimelbrant, A., J. N. Hutchinson, B. R. Thompson, *et al.*, 2007. Widespread monoallelic expression on human autosomes. Science, 318(5853): 1136–1140

Gregg, C., 2014. Known unknowns for allele-specific expression and genomic imprinting effects. F1000Prime Rep., 6: 75

Gregg, C., J. Zhang, J. E. Butler, *et al.*, 2010a. Sex-specific parent-of-origin allelic expression in the mouse brain. Science, 329(5992): 682–685

Gregg, C., J. Zhang, B. Weissbourd, *et al.*, 2010b. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science, 329(5992): 643–648

Groth, A., W. Rocha, A. Verreault, *et al.*, 2007. Chromatin challenges during DNA replication and repair. Cell, 128(4): 721–733

Hager, R., J. M. Cheverud, L. J. Leamy, *et al.*, 2008. Sex dependent imprinting effects on complex traits in mice. BMC Evol. Biol., 8: 303

Hager, R. and R. A. Johnstone, 2003. The genetic basis of family conflict resolution in mice. Nature, 421(6922): 533–535

Hager, R. and R. A. Johnstone, 2005. Differential growth of own and alien pups in mixed litters of mice: A role for genomic imprinting? Ethology, 111(8): 705–714

Haig, D., 1997. Parental antagonism, relatedness asymmetries, and genomic imprinting. Proc. Biol. Sci., 264(1388): 1657–1662

Haig, D., 2000. The kinship theory of genomic imprinting. Annu. Rev. Ecol. Syst., 31(1): 9–32

Haig, D. and C. Graham, 1991. Genomic imprinting and the strange case of the insulin-like growth factor II receptor. Cell, 64(6): 1045–1046

Haig, D. and M. Westoby, 1989. Parent-specific gene expression and the triploid endosperm. Am. Nat., 134(1): 147–155

Hall, J. G., 1990. Genomic imprinting: review and relevance to human diseases. Am. J. Hum. Genet., 46(5): 857–873

Hark, A. T., C. J. Schoenherr, D. J. Katz, *et al.*, 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19/Igf2* locus. Nature, 405(6785): 486–489

Hayward, B. E., A. Barlier, M. Korbonits, *et al.*, 2001. Imprinting of the $G_s\alpha$ gene *GNAS1* in the pathogenesis of acromegaly. J. Clin. Invest., 107(6): R31–R36

Holliday, R., 2006. Epigenetics: a historical overview. Epigenetics, 1(2): 76–80

Horsthemke, B., 2014. In brief: gGenomic imprinting and imprinting diseases. J. Pathol., 232(5): 485–487

Hou, S., Y. Chen, J. Liang, *et al.*, 2010. Developmental stage-specific imprinting of *IPL* in domestic pigs (*Sus scrofa*). J. Biomed. Biotechnol., 2010: 527–539

Hutter, B., M. Bieg, V. Helms, *et al.*, 2010. Divergence of imprinted genes during mammalian evolution. BMC Evol. Biol., 10: 116

Huynh, K. D. and J. T. Lee, 2003. Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. Nature, 426(6968): 857–862

Huynh, K. D. and J. T. Lee, 2005. X-chromosome inactivation: a hypothesis linking ontogeny and phylogeny. Nat. Rev. Genet., 6(5): 410–418

Ideraabdullah, F. Y., S. Vigneau, and M. S. Bartolomei, 2008. Genomic imprinting mechanisms in mammals. Mutat. Res., 647(1-2): 77–85

Illingworth, R. S. and A. P. Bird, 2009. CpG islands – "A rough guide". FEBS Lett., 583(11): 1713–1720

Iwasa, Y. and A. Pomiankowski, 1999. Sex specific X chromosome expression caused by genomic imprinting. J. Theor. Biol., 197(4): 487–495

Jablonka, E. and M. J. Lamb, 2002. The changing concept of epigenetics. Ann. N. Y. Acad. Sci., 981: 82–96

Jeltsch, A., 2002. Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA

methyltransferases. ChemBioChem, 3(4): 274–293

Jenuwein, T. and C. D. Allis, 2001. Translating the histone code. Science, 293(5532): 1074–1080

Jiang, Y. H., J. Bressler, and A. L. Beaudet, 2004. Epigenetics and human disease. Annu. Rev. Genomics Hum. Genet., 5: 479–510

Jinno, Y., K. Sengoku, M. Nakao, *et al.*, 1996. Mouse/human sequence divergence in a region with a paternal-specific methylation imprint at the human *H19* locus. Hum. Mol. Genet., 5(8): 1155–1161

Jones, P. A. and S. B. Baylin, 2007. The epigenomics of cancer. Cell, 128(4): 683–692

Jones, P. L., G. J. Veenstra, P. A. Wade, *et al.*, 1998. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat. Genet., 19(2): 187–191

Kaffer, C. R., A. Grinberg, and K. Pfeifer, 2001. Regulatory mechanisms at the mouse *Igf2/H19* locus. Mol. Cell Biol., 21(23): 8189–8196

Kaffer, C. R., M. Srivastava, K. Y. Park, *et al.*, 2000. A transcriptional insulator at the imprinted *H19/Igf2* locus. Genes Dev., 14(15): 1908–1919

Kafri, T., M. Ariel, M. Brandeis, *et al.*, 1992. Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes Dev., 6(5): 705–714

Kaikkonen, M. U., M. T. Lam, and C. K. Glass, 2011. Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc. Res., 90(3): 430–440

Kanduri, C., V. Pant, D. Loukinov, *et al.*, 2000. Functional association of CTCF with the insulator upstream of the *H19* gene is parent of origin-specific and methylation-sensitive. Curr. Biol., 10(14): 853–856

Kelsey, G. and M. S. Bartolomei, 2012. Imprinted genes ... and the number is? PLoS Genet., 8(3): e1002601

Kermicle, J. L., 1970. Dependence of the R-mottled aleurone phenotype in maize on mode of sexual transmission. Genetics, 66(1): 69–85

Khatib, H., 2007. Is it genomic imprinting or preferential expression? Bioessays, 29(10): 1022–1028

Khosla, S., G. Mendiratta, and V. Brahmachari, 2006. Genomic imprinting in the mealybugs. Cytogenet. Genome Res., 113(1-4): 41–52

Kim, J., A. Kollhoff, A. Bergmann, *et al.*, 2003. Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, *Peg3*. Hum. Mol. Genet., 12(3): 233–245

Kouzarides, T., 2007. Chromatin modifications and their function. Cell, 128(4): 693–705

Krause, B., L. Sobrevia, and P. Casanello, 2009. Epigenetics: new concepts of old phenomena in

vascular physiology. Curr. Vasc. Pharmacol., 7(4): 513–520

Kuo, M. H. and C. D. Allis, 1998. Roles of histone acetyltransferases and deacetylases in gene regulation. Bioessays, 20(8): 615–626

Lander, E. S., L. M. Linton, B. Birren, *et al.*, 2001. Initial sequencing and analysis of the human genome. Nature, 409(6822): 860–921

Lawson, H. A., J. M. Cheverud, and J. B. Wolf, 2013. Genomic imprinting and parent-of-origin effects on complex traits. Nat. Rev. Genet., 14(9): 609–617

Li, L., E. B. Keverne, S. A. Aparicio, *et al.*, 1999. Regulation of maternal behavior and offspring growth by paternally expressed *Peg3*. Science, 284(5412): 330–333

Li, S. M., Z. Valo, J. Wang, *et al.*, 2012. Transcriptome-wide survey of mouse CNS-derived cells reveals monoallelic expression within novel gene families. PLoS ONE, 7(2): e31751

Lister, R., R. C. O'Malley, J. Tonti-Filippini, *et al.*, 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell, 133(3): 523–536

Lister, R., M. Pelizzola, R. H. Dowen, *et al.*, 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462(7271): 315–322

Lloyd, V., 2000. Parental imprinting in *Drosophila*. Genetica, 109(1-2): 35–44

Luedi, P. P., A. J. Hartemink, and R. L. Jirtle, 2005. Genome-wide prediction of imprinted murine genes. Genome Res., 15(6): 875–884

Luger, K., A. W. Mader, R. K. Richmond, *et al.*, 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature, 389(6648): 251–260

MacDonald, M., T. Hassold, J. Harvey, *et al.*, 1994. The origin of 47,XXY and 47,XXX aneuploidy: heterogeneous mechanisms and role of aberrant recombination. Hum. Mol. Genet., 3(8): 1365–1371

Mak, W., T. B. Nesterova, M. de Napoles, *et al.*, 2004. Reactivation of the paternal X chromosome in early mouse embryos. Science, 303(5658): 666–669

Mancini-Dinardo, D., S. J. Steele, J. M. Levorse, *et al.*, 2006. Elongation of the *Kcnq1ot1* transcript is required for genomic imprinting of neighboring genes. Genes Dev., 20(10): 1268–1282

Mantovani, G., E. Ballare, E. Giammona, *et al.*, 2002. The Gsα gene: predominant maternal origin of transcription in human thyroid gland and gonads. J. Clin. Endocrinol. Metab., 87(10): 4736–4740

Matsuo, K., O. Clay, T. Takahashi, *et al.*, 1993. Evidence for erosion of mouse CpG islands during mammalian evolution. Somat. Cell Mol. Genet., 19(6): 543–555

Matzke, M. and A. J. M. Matzke, 1993. Genomic imprinting in plants: parental effects and trans-inactivation phenomena. Annu. Rev. Plant Physiol. Plant Mol. Biol., 44(1): 53–76

McEwen, K. R. and A. C. Ferguson-Smith, 2009. Genomic imprinting – a model for roles of histone modifications in epigenetic control. In A. C. Ferguson-Smith, J. M. Greally, and R. A. Martienssen (Editors), *Epigenomics*, pp. 235–258. Springer Netherlands

McGowan, R. A. and C. C. Martin, 1997. DNA methylation and genome imprinting in the zebrafish, *Danio rerio*: some evolutionary ramifications. Biochem. Cell Biol., 75(5): 499–506

McGrath, J. and D. Solter, 1984. Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell, 37(1): 179–183

McVean, G. T. and L. D. Hurst, 1997. Molecular evolution of imprinted genes: no evidence for antagonistic coevolution. Proc. Biol. Sci., 264(1382): 739–746

Meijers-Heijboer, E. J., L. A. Sandkuijl, H. G. Brunner, *et al.*, 1992. Linkage analysis with chromosome 15q11-13 markers shows genomic imprinting in familial Angelman syndrome. J. Med. Genet., 29(12): 853–857

Meissner, A., A. Gnirke, G. W. Bell, *et al.*, 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res., 33(18): 5868–5877

Menheniott, T. R., K. Woodfine, R. Schulz, *et al.*, 2008. Genomic imprinting of Dopa decarboxylase in heart and reciprocal allelic expression with neighboring *Grb10*. Mol. Cell Biol., 28(1): 386–396

Miyake, T., N. Takebayashi, and D. E. Wolf, 2009. Possible diversifying selection in the imprinted gene, *MEDEA*, in *Arabidopsis*. Mol. Biol. Evol., 26(4): 843–857

Monk, M. and M. Grant, 1990. Preferential X-chromosome inactivation, DNA methylation and imprinting. Development (Suppl.), pp. 55–62

Moore, L. D., T. Le, and G. Fan, 2013. DNA methylation and its basic function. Neuropsychopharmacology, 38(1): 23–38

Moore, T. and D. Haig, 1991. Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet., 7(2): 45–49

Moore, T. and W. Mills, 2008. Evolutionary theories of imprinting – enough already! Adv. Exp. Med. Biol., 626: 116–122

Morcos, L., B. Ge, V. Koka, *et al.*, 2011. Genome-wide assessment of imprinted expression in human cells. Genome Biol., 12(3): R25

Morison, I. M., J. P. Ramsay, and H. G. Spencer, 2005. A census of mammalian imprinting. Trends Genet., 21(8): 457–465

Murrell, A., P. J. Hurd, and I. C. Wood, 2013. Epigenetic mechanisms in development and disease. Biochem. Soc. Trans., 41(3): 697–699

Nabetani, A., I. Hatada, H. Morisaki, *et al.*, 1997. Mouse *U2af1-rs1* is a neomorphic imprinted gene. Mol. Cell Biol., 17(2): 789–798

Nan, X., H. H. Ng, C. A. Johnson, *et al.*, 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature, 393(6683): 386–389

Netchine, I., S. Rossignol, M. N. Dufourg, *et al.*, 2007. 11p15 imprinting center region 1 loss of methylation is a common and specific cause of typical Russell-Silver syndrome: clinical scoring system and epigenetic-phenotypic correlations. J. Clin. Endocrinol. Metab., 92(8): 3148–3154

Nicholls, R. D., S. Saitoh, and B. Horsthemke, 1998. Imprinting in Prader-Willi and Angelman syndromes. Trends Genet., 14(5): 194–200

O'Connell, M. J., N. B. Loughran, T. A. Walsh, *et al.*, 2010. A phylogenetic approach to test for evidence of parental conflict or gene duplications associated with protein-encoding imprinted orthologous genes in placental mammals. Mamm. Genome, 21(9-10): 486–498

Ohno, S., W. D. Kaplan, and R. Kinosita, 1959. Formation of the sex chromatin by a single X-chromosome in liver cells of *Rattus norvegicus*. Exp. Cell Res., 18: 415–418

Okamoto, I., A. P. Otte, C. D. Allis, *et al.*, 2004. Epigenetic dynamics of imprinted X inactivation during early mouse development. Science, 303(5658): 644–649

O'Neill, M. J., R. S. Ingram, P. B. Vrana, *et al.*, 2000. Allelic expression of *IGF2* in marsupials and birds. Dev. Genes Evol., 210(1): 18–20

Parker-Katiraee, L., A. R. Carson, T. Yamada, *et al.*, 2007. Identification of the imprinted *KLF14* transcription factor undergoing human-specific accelerated evolution. PLoS Genet., 3(5): e65

Patten, M. M., L. Ross, J. P. Curley, *et al.*, 2014. The evolution of genomic imprinting: theories, predictions and empirical tests. Heredity (Edinb), 113(2): 119–128

Pembrey, M., 2012. *An introduction to the Genetics and Epigenetics of Human Disease.* Progress Educational Trust, London, UK

Penaherrera, M. S., S. Weindler, M. I. Van Allen, *et al.*, 2010. Methylation profiling in individuals with Russell-Silver syndrome. Am. J. Med. Genet. A, 152A(2): 347–355

Peters, J., 2014. The role of genomic imprinting in biology and disease: an expanding view. Nat. Rev. Genet., 15(8): 517–530

Peters, J. and C. M. Williamson, 2007. Control of imprinting at the *Gnas* cluster. Epigenetics, 2(4): 207–213

Pfeifer, K., 2000. Mechanisms of genomic imprinting. Am. J. Hum. Genet., 67(4): 777–787

Phillips, J. E. and V. G. Corces, 2009. CTCF: master weaver of the genome. Cell, 137(7): 1194–1211

Pidsley, R., C. Fernandes, J. Viana, *et al.*, 2012. DNA methylation at the *Igf2/H19* imprinting control region is associated with cerebellum mass in outbred mice. Mol. Brain, 5: 42

Ponting, C. P., P. L. Oliver, and W. Reik, 2009. Evolution and functions of long noncoding RNAs. Cell, 136(4): 629–641

Prickett, A. R. and R. J. Oakey, 2012. A survey of tissue-specific genomic imprinting in mammals. Mol. Genet. Genomics, 287(8): 621–630

Ptashne, M., 2007. On the use of the word 'epigenetic'. Curr. Biol., 17(7): R233–236

Qian, N., D. Frank, D. O'Keefe, *et al.*, 1997. The *IPL* gene on chromosome 11p15.5 is imprinted in humans and mice and is similar to *TDAG51*, implicated in Fas expression and apoptosis. Hum. Mol. Genet., 6(12): 2021–2029

Qiu, J., 2006. Epigenetics: unfinished symphony. Nature, 441(7090): 143–145

Raefski, A. S. and M. J. O'Neill, 2005. Identification of a cluster of X-linked imprinted genes in mice. Nat. Genet., 37(6): 620–624

Ratel, D., J. L. Ravanat, F. Berger, *et al.*, 2006. N6-methyladenine: the other methylated base of DNA. Bioessays, 28(3): 309–315

Razin, A. and H. Cedar, 1991. DNA methylation and gene expression. Microbiol. Rev., 55(3): 451–458

Razin, A. and H. Cedar, 1994. DNA methylation and genomic imprinting. Cell, 77(4): 473–476

Reik, W. and J. Walter, 2001. Genomic imprinting: parental influence on the genome. Nat. Rev. Genet., 2(1): 21–32

Renfree, M. B., E. I. Ager, G. Shaw, *et al.*, 2008. Genomic imprinting in marsupial placentation. Reproduction, 136(5): 523–531

Riggs, A. D., R. A. Martienssen, and V. E. A. Russo, 1996. Introduction. In V. E. A. Russo, R. A. Martienssen, and A. D. Riggs (Editors), *Epigenetic Mechanisms of Gene Regulation*, pp. 1–4. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Riggs, A. D. and T. N. Porter, 1996. Overview of epigenetic mechanisms. In V. E. A. Russo, R. A. Martienssen, and A. D. Riggs (Editors), *Epigenetic Mechanisms of Gene Regulation*, pp. 29–45. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Rivera, R. M. and L. B. Bennett, 2010. Epigenetics in humans: an overview. Curr. Opin. Endocrinol. Diabetes Obes., 17(6): 493–499

Robertson, K. D., 2005. DNA methylation and human disease. Nat. Rev. Genet., 6(8): 597–610

Routh, A., S. Sandin, and D. Rhodes, 2008. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. Proc. Natl. Acad. Sci. U.S.A., 105(26): 8872–8877

Russell, A., 1954. A syndrome of intra-uterine dwarfism recognizable at birth with cranio-facial dysostosis, disproportionately short arms, and other anomalies (5 examples). Proc. R. Soc. Med., 47(12): 1040–1044

Ruthenburg, A. J., H. Li, D. J. Patel, *et al.*, 2007. Multivalent engagement of chromatin modifications by linked binding modules. Nat. Rev. Mol. Cell Biol., 8(12): 983–994

Samollow, P. B., E. S. Robinson, A. L. Ford, *et al.*, 1995. Developmental progression of *Gpd* expression from the inactive X chromosome of the Virginia opossum. Dev. Genet., 16(4): 367–378

Schulz, R., K. Woodfine, T. R. Menheniott, *et al.*, 2008. WAMIDEX: a web atlas of murine genomic imprinting and differential expression. Epigenetics, 3(2): 89–96

Scott, R. J., M. Spielman, J. Bailey, *et al.*, 1998. Parent-of-origin effects on seed development in *Arabidopsis thaliana*. Development, 125(17): 3329–3341

Selinger, D. A. and V. L. Chandler, 2001. *B-Bolivia*, an allele of the maize *b1* gene with variable expression, contains a high copy retrotransposon-related sequence immediately upstream. Plant Physiol., 125(3): 1363–1379

Sha, K. and A. Fire, 2005. Imprinting capacity of gamete lineages in Caenorhabditis elegans. Genetics, 170(4): 1633–1652

Shirane, K., H. Toh, H. Kobayashi, *et al.*, 2013. Mouse oocyte methylomes at base resolution reveal genome-wide accumulation of non-CpG methylation and role of DNA methyltransferases. PLoS Genet., 9(4): e1003439

Silver, H. K., W. Kiyasu, J. George, *et al.*, 1953. Syndrome of congenital hemihypertrophy, shortness of stature, and elevated urinary gonadotropins. Pediatrics, 12(4): 368–376

Smith, F. M., L. J. Holt, A. S. Garfield, *et al.*, 2007. Mice with a disruption of the imprinted *Grb10* gene exhibit altered body composition, glucose homeostasis, and insulin signaling during postnatal life. Mol. Cell Biol., 27(16): 5871–5886

Smith, N. G. and L. D. Hurst, 1998. Molecular evolution of an imprinted gene: repeatability of patterns of evolution within the mammalian insulin-like growth factor type II receptor. Genetics, 150(2): 823–833

Smits, G., A. J. Mungall, S. Griffiths-Jones, *et al.*, 2008. Conservation of the *H19* noncoding RNA and *H19-IGF2* imprinting mechanism in therians. Nat. Genet., 40(8): 971–976

Solter, D., 1992. Relevance of genomic imprinting to human diseases. Curr. Opin. Biotechnol., 3(6): 632–636

Spencer, H. G., 2009. Effects of genomic imprinting on quantitative traits. Genetica, 136(2): 285–293

Spencer, H. G. and A. G. Clark, 2014. Non-conflict theories for the evolution of genomic imprinting. Heredity (Edinb), 113(2): 112–118

Spencer, H. G., M. W. Feldman, and A. G. Clark, 1998. Genetic conflicts, multiple paternity and the evolution of genomic imprinting. Genetics, 148(2): 893–904

Spillane, C., K. J. Schmid, S. Laoueille-Duprat, *et al.*, 2007. Positive darwinian selection at the imprinted *MEDEA* locus in plants. Nature, 448(7151): 349–352

Strunnikova, M., U. Schagdarsurengin, A. Kehlen, *et al.*, 2005. Chromatin inactivation precedes de novo DNA methylation during the progressive epigenetic silencing of the *RASSF1A* promoter. Mol. Cell Biol., 25(10): 3923–3933

Suzuki, S., M. B. Renfree, A. J. Pask, *et al.*, 2005. Genomic imprinting of *IGF2*, $p57^{KIP2}$ and *PEG1/MEST* in a marsupial, the tammar wallaby. Mech. Dev., 122(2): 213–222

Szabó, P. E., S.-H. E. Tang, A. Rentsendorj, *et al.*, 2000. Maternal-specific footprints at putative CTCF sites in the *H19* imprinting control region give evidence for insulator function. Curr. Biol., 10(10): 607–610

Szabó, P. E., S.-H. E. Tang, F. J. Silva, *et al.*, 2004. Role of CTCF binding sites in the *Igf2/H19* imprinting control region. Mol. Cell Biol., 24(11): 4791–4800

Tada, T., M. Tada, K. Hilton, *et al.*, 1998. Epigenotype switching of imprintable loci in embryonic germ cells. Dev. Genes Evol., 207(8): 551–561

Takagi, N. and M. Sasaki, 1975. Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. Nature, 256(5519): 640–642

Tang, F., C. Barbacioru, E. Nordman, *et al.*, 2011. Deterministic and stochastic allele specific gene expression in single mouse blastomeres. PLoS ONE, 6(6): e21208

The ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science, 306(5696): 636–640

The ENCODE Project Consortium, 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447(7146): 799–816

Thorvaldsen, J. L. and M. S. Bartolomei, 2007. SnapShot: imprinted gene clusters. Cell, 130(5): 958

Tollefsbol, T. (Editor), 2012. *Epigenetics in Human Disease.* Academic Press, MA, USA

Tremblay, K. D., K. L. Duran, and M. S. Bartolomei, 1997. A 5' 2-kilobase-pair region of the imprinted mouse *H19* gene exhibits exclusive paternal methylation throughout development. Mol. Cell Biol., 17(8): 4322–4329

Vanyushin, B. F., 2006. DNA methylation in plants. Curr. Top. Microbiol. Immunol., 301: 67–122

Vasques, L. R., M. N. Klockner, and L. V. Pereira, 2002. X chromosome inactivation: how human are mice? Cytogenet. Genome Res., 99(1-4): 30–35

Verona, R. I., M. R. Mann, and M. S. Bartolomei, 2003. Genomic imprinting: intricacies of epigenetic regulation in clusters. Annu. Rev. Cell Dev. Biol., 19: 237–259

Vinkenoog, R., C. Bushell, M. Spielman, *et al.*, 2003. Genomic imprinting and endosperm development in flowering plants. Mol. Biotechnol., 25(2): 149–184

Virani, S., J. A. Colacino, J. H. Kim, *et al.*, 2012. Cancer epigenetics: a brief review. ILAR J,

53(3-4): 359–369

Waddington, C. H., 1939. Preliminary Notes on the Development of the Wings in Normal and Mutant Strains of *Drosophila*. Proc. Natl. Acad. Sci. U.S.A., 25(7): 299–307

Waddington, C. H., 2012. The epigenotype. Int. J. Epidemiol., 41(1): 10–13

Wagschal, A. and R. Feil, 2006. Genomic imprinting in the placenta. Cytogenet. Genome Res., 113(1-4): 90–98

Wan, L. B. and M. S. Bartolomei, 2008. Regulation of imprinting in clusters: noncoding RNAs versus insulators. Adv. Genet., 61: 207–223

Wang, Y., K. Joh, S. Masuko, *et al.*, 2004. The mouse *Murr1* gene is imprinted in the adult brain, presumably due to transcriptional interference by the antisense-oriented *U2af1-rs1* gene. Mol. Cell Biol., 24(1): 270–279

Weber, M., I. Hellmann, M. B. Stadler, *et al.*, 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet., 39(4): 457–466

West, J. D., W. I. Frels, V. M. Chapman, *et al.*, 1977. Preferential expression of the maternally derived X chromosome in the mouse yolk sac. Cell, 12(4): 873–882

West, J. D., V. E. Papaioannou, W. I. Frels, *et al.*, 1978. Preferential expression of the maternally derived X chromosome in extraembryonic tissues of the mouse. Basic Life Sci., 12: 361–377

Wilkins, J. F. and F. Ubeda, 2011. Diseases associated with genomic imprinting. Prog. Mol. Biol. Transl. Sci., 101: 401–445

Williamson, C. M., S. T. Ball, W. T. Nottingham, *et al.*, 2004. A *cis*-acting control region is required exclusively for the tissue-specific imprinting of *Gnas*. Nat. Genet., 36(8): 894–899

Williamson, C. M., A. Blake, S. Thomas, *et al.*, 2013. Mouse Imprinting Data and References. MRC Harwell, Oxfordshire. World Wide Web Site http://www.har.mrc.ac.uk/research/genomic_imprinting/

Wolf, J. B. and R. Hager, 2006. A maternal-offspring coadaptation theory for the evolution of genomic imprinting. PLoS Biol., 4(12): e380

Wolf, J. B., R. Hager, and J. M. Cheverud, 2008. Genomic imprinting effects on complex traits: a phenotype-based perspective. Epigenetics, 3(6): 295–299

Wolff, P., I. Weinhofer, J. Seguin, *et al.*, 2011. High-resolution analysis of parent-of-origin allelic expression in the *Arabidopsis* endosperm. PLoS Genet., 7(6): e1002126

Wollmann, H. A., T. Kirchner, H. Enders, *et al.*, 1995. Growth and symptoms in Silver-Russell syndrome: review on the basis of 386 patients. Eur. J. Pediatr., 154(12): 958–968

Wood, A. J. and R. J. Oakey, 2006. Genomic imprinting in mammals: emerging themes and established theories. PLoS Genet., 2(11): e147

Xie, W., C. L. Barr, A. Kim, *et al.*, 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell, 148(4): 816–831

Xue, F., X. C. Tian, F. Du, *et al.*, 2002. Aberrant patterns of X chromosome inactivation in bovine clones. Nat. Genet., 31(2): 216–220

Yan, J., J. R. Zierath, and R. Barres, 2011. Evidence for non-CpG methylation in mammals. Exp. Cell Res., 317(18): 2555–2561

Yang, T., T. E. Adamson, J. L. Resnick, *et al.*, 1998. A mouse model for Prader-Willi syndrome imprinting-centre mutations. Nat. Genet., 19(1): 25–31

Yu, S., D. Yu, E. Lee, *et al.*, 1998. Variable and tissue-specific hormone resistance in heterotrimeric G_s protein α-subunit ($G_s\alpha$) knockout mice is due to tissue-specific imprinting of the $G_s\alpha$ gene. Proc. Natl. Acad. Sci. U.S.A., 95(15): 8715–8720

Zhou, H., H. Hu, and M. Lai, 2010. Non-coding RNAs and their epigenetic regulatory mechanisms. Biol. Cell, 102(12): 645–655

Zwemer, L. M., A. Zak, B. R. Thompson, *et al.*, 2012. Autosomal monoallelic expression in the mouse. Genome Biol., 13(2): R10

Chapter 3

Analysis of Imprinting under a Quantitative Genetics Framework: an Overview

Due to a potential impact of genomic imprinting on complex traits, as discussed in Chapter 2, and the fact that imprinting is transmissible across generations, studies focusing on the contribution of genomic imprinting to phenotypic variation are of interest. Models incorporating imprinting effects have been proposed in a quantitative genetics context. In this chapter, one of the models is introduced and some applications of this model are provided.

3.1 A One-locus Quantitative Imprinting Model

Given a biallelic locus, there are four possible genotypes: A_1A_1, A_1A_2, A_2A_1 and A_2A_2 (maternally inherited allele written first, throughout the entire document). In standard quantitative genetic model where imprinting is not considered, genotypes A_1A_2 and A_2A_1 are not differentiated and hence have the same genotypic value d (Falconer and Mackay, 1996; Lynch and Walsh, 1998). Since

genomic imprinting is featured by mono-allelic expression depending on the parent-of-origin from which this allele is inherited, paternal and maternal alleles would be functionally nonequivalent. Therefore, genotypes A_1A_1 and A_2A_1 will be phenotypically identical for maternal imprinting since only the paternal allele (A_1) is active, resulting in a combined genotype $\_A_1$. Similarly, A_1A_2 and A_2A_2 reduce to $\_A_2$ and hence only two functional genotypes are observable given an imprinting direction. More generally, expression repression of imprinted allele may not be complete; this phenomenon is called partial, as opposed to complete, imprinting (Khatib, 2007; Wolf *et al.*, 2008a; Morcos *et al.*, 2011), indicating that the two reciprocal heterozygotes A_1A_2 and A_2A_1 may have different genotypic values. Therefore, all four genotypes have unique and identifiable genotypic values (Figure 3.1), as proposed by Spencer (2002), Shete and Amos (2002), and de Koning *et al.* (2002). In this model, genotypic values of A_1A_1 and A_2A_2 are same as in the standard model, but genotypic values d_1 and d_2 are assigned to A_1A_2 and A_2A_1, respectively (Spencer, 2002). Equivalently, Shete and Amos (2002) and de Koning *et al.* (2002) used parameterization $d_1 = d - i$ and $d_2 = d + i$ to described the genotypic values of the two heterozygotes. This parameterization kept the conventional dominance effect d and defined i as the imprinting effect. Because i is isolated in this parameterization, impact of imprinting on genotypic values, allelic substitution effect, breeding values, and variance decomposition becomes more intuitive and easier to interpret. Therefore, this parameterization is adopted throughout the entire document.

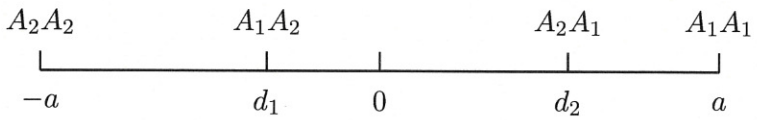


Figure 3.1: Genotypic values of the four possible genotypes in a biallelic imprinted locus (Spencer, 2002).

3.1.1 Population Mean and Breeding Values

Under Hardy-Weinberg assumptions, the genotypic frequencies of these four genotypes are

$$\Pr(A_1A_1) = p^2,$$
$$\Pr(A_2A_1) = qp,$$
$$\Pr(A_1A_2) = pq,$$
$$\Pr(A_2A_2) = q^2,$$

where p and q $(p+q=1)$ are allelic frequencies of the A_1 and A_2 allele, respectively. Given this Hardy-Weinberg frequency and the genotypic values defined in Figure 3.1, the population mean is

$$\mu = p^2(a) + qp(d+i) + pq(d-i) + q^2(-a) = a(q-p) + 2pqd.$$

Note that this value is the same as in a standard additive model where imprinting is not considered. The next step in Spencer (2002) was to derive the breeding value of different genotypes by taking the difference between the population mean and the mean genotypic value of the offspring of a genotype when it was mated to an individual randomly drawn from the population. Because each genotype can act as a sire or a dam under imprinting, there would be eight distinct breeding values (Table 3.1). Alternatively, breeding values can be obtained via use of the allelic substitution effect, which is defined in a regression model as shown in Lynch and Walsh (1998). When applied to imprinting, Shete and Amos (2002) adapted the model as

$$G_{kl} = \mu + I_{♀}\alpha_{♀} + I_{♂}\alpha_{♂} + \delta_{kl}, \tag{3.1}$$

where $G_{kl} = \{-a, d-i, d+i, a\}$ is the genotypic value of the A_kA_l genotype; $I_{♀}$ and $I_{♂}$ are (0, 1) indicator variables denoting the number of A_1 allele inherited from a specific parent (e.g., $I_{♀} = 1$ and $I_{♂} = 0$ describes that the genotype is A_1A_2); and δ_{kl} is the model residual, interpreted as dominance deviation in Lynch and Walsh (1998). $\alpha_{♀}$ and $\alpha_{♂}$ are regression coefficients associated with $I_{♀}$ and $I_{♂}$, and are defined as maternal and paternal allelic substitution effects, respectively. Both Shete

and Amos (2002) and Spencer (2002) arrived at $\alpha_{♀} = a+(q-p)d-i$ and $\alpha_{♂} = a+(q-p)d+i$, and it was observed that the coefficients that convert allelic substitution effects into breeding values under imprinting are the same as those in the standard model (Table 3.1). Note that $(\alpha_{♂} - \alpha_{♀})/2 = i$, the imprinting effect, and $(\alpha_{♂} + \alpha_{♀})/2 = \alpha$, the average allelic substitution effect in the standard sense. de Koning *et al.* (2002) presented the same result.

Table 3.1: Breeding values of all four genotypes in two sexes under genomic imprinting (Spencer, 2002).

	As female		As male	
Genotype	$Freq.^1$	$BV_{kl♀}$	$Freq.^1$	$BV_{kl♂}$
A_1A_1	$p^2/2$	$2q(a+(q-p)d-i)$	$p^2/2$	$2q(a+(q-p)d+i)$
A_2A_1	$qp/2$	$(q-p)(a+(q-p)d-i)$	$qp/2$	$(q-p)(a+(q-p)d+i)$
A_1A_2	$pq/2$	$(q-p)(a+(q-p)d-i)$	$pq/2$	$(q-p)(a+(q-p)d+i)$
A_2A_2	$q^2/2$	$-2p(a+(q-p)d-i)$	$q^2/2$	$-2p(a+(q-p)d+i)$

[1] Values are frequencies of a genotype in the entire population, given a sex. Here a 1:1 sex ratio is assumed.

3.1.2 Decomposition of Genetic Variance

With four genotypic values and the associated frequencies, it is easy to find that

$$\begin{aligned} E(G_{kl}^2) &= p^2a^2 + qp(d+i)^2 + pq(d-i)^2 + q^2(-a)^2 \\ &= (p^2+q^2)a^2 + 2pq(d^2+i^2), \end{aligned}$$

which gives the genetic variance

$$\sigma_G^2 = E(G_{kl}^2) - \mu^2 = 2pq(a+d(q-p))^2 + (2pqd)^2 + 2pqi^2. \tag{3.2}$$

Note that the first two terms, $2pq(a+d(q-p))^2 + (2pqd)^2$, give the genetic variance assuming additivity and dominance, indicating that genetic variance increases by $2pqi^2$ under imprinting. The additive genetic variance can be obtained by taking the variance of the eight breeding values

in Table 3.1, giving that

$$\sigma_A^2 = 2pq(a + d(q - p))^2 + 2pqi^2. \tag{3.3}$$

Alternatively, Shete and Amos (2002) defined additive genetic variance within females and males as $\sigma^2_{A♀} = pq\alpha^2_{♀}$ and $\sigma^2_{A♂} = pq\alpha^2_{♂}$ using the maternal and paternal allelic substitution effects respectively, and the same result as Equation 3.3 was obtained by summing $\sigma^2_{A♀}$ and $\sigma^2_{A♂}$ under a 1:1 sex ratio assumption. By comparing Equations 3.2 and 3.3, one may deduce that the increase of genetic variation is entirely due to the impact of genomic imprinting on additive genetic variance.

3.2 Mapping Imprinted QTL

In quantitative genetics, finding the position of genomic regions with large effects on a target trait is of interest. This procedure is known as QTL (quantitative trait loci) mapping. Molecular markers play an important role in QTL mapping studies and various statistical methods have been developed since the 1980's. In this section, basic methods of QTL mapping are briefly introduced, followed by a more detailed description of the development of methods for mapping putatively imprinted QTL (iQTL) and some applications.

3.2.1 QTL Mapping Basics

With a segregating marker at a given locus, the simplest strategy for detecting QTL is probably checking the phenotypic contrasts between different marker genotype groups with a t-test (for backcross design) or an F-test (for intercross design) under normality assumptions. This is also known as single marker regression, where different marker genotype groups are regarded as different values of a factorized regression covariate (Soller *et al.*, 1976). Evidence of the existence of a QTL is indicated by a LOD score, typically calculated as

$$\text{LOD} = \frac{n}{2} \log_{10} \left(\frac{RSS_0}{RSS_1} \right), \tag{3.4}$$

where n is the number of observations in the data set, RSS_0 and RSS_1 are residual sum of squares of the regression models under the null (no QTL) and alternative (QTL exists) hypotheses, respectively. Marker regression for QTL detection is simple and easy to implement. However, even if a QTL may exist in the vicinity of the given marker, as indicated by a large LOD score, precise inference on the location of the putative QTL is usually difficult.

QTL detection studies did not become QTL "mapping" until the advent of interval mapping (Lander and Botstein, 1989), where the location of a putative QTL can be "mapped" at a locus of a genetic map using a pair of flanking markers. Consider again the null hypothesis that there is no QTL anywhere in the genome, i.e., $H_0 : y_i \sim N(\mu_0, \sigma^2)$ under a normality assumption. The likelihood function under H_0 is

$$\mathcal{L}(\mu_0, \sigma^2) = \prod_i N(y_i|\mu_0, \sigma^2).$$

For the alternative hypothesis H_1, suppose there is a QTL with distinct genotypes QQ and Qq in a backcross design, the phenotypic distribution under H_1 is then a mixture of normal densities:

$$\sum_j p_{ij} N(y_i|\mu_j, \sigma^2), \tag{3.5}$$

where p_{ij} is the probability of observing a genotype j ($= QQ$ or Qq) at the given locus in individual i, and μ_j is the expected phenotypic value for genotype j. Note here that a common dispersion parameter σ^2 is assumed across different genotypes. The likelihood function under H_1 is expressed as

$$\mathcal{L}(\boldsymbol{\mu}, \sigma^2) = \prod_i \sum_j p_{ij} N(y_i|\mu_j, \sigma^2), \tag{3.6}$$

where $\boldsymbol{\mu}$ is a vector containing the expected phenotypes of all possible QTL genotypes. After obtaining an estimate of μ_0, $\boldsymbol{\mu}$, and σ^2 (usually via maximum likelihood), the LOD score is calculated

as (assuming p_{ij} is known)

$$\text{LOD} = \log_{10}\left(\frac{\mathcal{L}(\hat{\mu}_0, \hat{\sigma}^2)}{\mathcal{L}(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2)}\right). \tag{3.7}$$

Because the QTL genotype is not observable, p_{ij} is usually obtained in the following way. Suppose two markers are located on both sides of an unknown QTL, denoted as M_l and M_r, and the genetic distance between M_l and M_r (Δ_{lr}) is known. In a backcross design, all F_1 individuals are with M_lM_r/m_lm_r marker genotypes, where haplotype M_lM_r is from one parent and m_lm_r is from the other. When crossing F_1 with one of the two parental lines, say M_lM_r/M_lM_r, there are four possible marker genotypes in total in the F_2 generation: M_lM_r/M_lM_r, M_lM_r/M_lm_r, M_lM_r/m_lM_r, and M_lM_r/m_lm_r. Considering two possible genotypes (QQ and Qq) at the QTL, there could be eight three-locus genotypes, where M_lQM_r/M_lQM_r and M_lQM_r/m_lqm_r are nonrecombinants and all others are recombinants. Because Δ_{lr} is known, the recombination rate r between M_l and QTL, QTL and M_r, and M_l and M_r can be calculated using say, Haldane's mapping function (Haldane, 1919), given a putative QTL location. With the r's, the probabilities of observing a certain QTL genotype given the marker genotype can be calculated (Table 3.2).

Table 3.2: Conditional probabilities for the QTL genotypes in a backcross design, given the genotypes at two flanking markers (bold faced = nonrecombinant genotypes).

Marker genotype		p_{ij} = Pr(QTL genotype | marker genotype)	
Left	Right	QQ	Qq
M_lM_l	M_rM_r	$\mathbf{(1-r_{lQ})(1-r_{rQ})/(1-r_{lr})}$	$r_{lQ}r_{rQ}/(1-r_{lr})$
M_lM_l	M_rm_r	$(1-r_{lQ})r_{rQ}/r_{lr}$	$r_{lQ}(1-r_{rQ})/r_{lr}$
M_lm_l	M_rM_r	$r_{lQ}(1-r_{rQ})/r_{lr}$	$(1-r_{lQ})r_{rQ}/r_{lr}$
M_lm_l	M_rm_r	$r_{lQ}r_{rQ}/(1-r_{lr})$	$\mathbf{(1-r_{lQ})(1-r_{rQ})/(1-r_{lr})}$

With this method, it is possible to obtain the probabilities of observing a QQ or a Qq QTL at any location between a pair of markers with known distance. Therefore, every position within the interval defined by these markers can be viewed as a putative QTL and a LOD score is obtained for each position. The position with the highest LOD score is then the most likely position of a QTL. Interval mapping has an attractive advantage that it can detect the existence of a QTL, and

also infer the location of the putative QTL. However, interval mapping is more computationally complicated. This is because the likelihood function $\mathcal{L}_1$ (Equation 3.6) is not in a closed form such that iterative algorithms, e.g., the expectation-maximization (EM) algorithm, are usually needed to obtain the maximum likelihood estimates of $\boldsymbol{\mu}$ and σ^2.

To overcome the computational difficulties in the standard interval mapping, note that $E(y_i|\boldsymbol{M}_i) = \sum_j p_{ij}\mu_j$ (Equation 3.5), indicating that the conditional expectation of the phenotypic value is linear in the μ_j's, given the available marker data, denoted by $\boldsymbol{M}_i$. Assuming that the phenotypic variance within each QTL genotype is the same, $y_i|\boldsymbol{M}_i$ can then be approximated by $N(\sum_j p_{ij}\mu_j, \sigma^2)$. This approximation makes it sufficient to evaluate the alternative hypothesis through the linear model

$$y_i = m + \sum_j p_{ij}\beta_j + e_i, \tag{3.8}$$

where m is a value common to all individuals; p_{ij} is the same as in standard interval mapping (Table 3.2) with associated regression coefficient β_j and the summation is over all possible QTL genotypes; e_i is the model residual with assumed distribution $N(0, \sigma^2)$. This method was proposed by Haley and Knott (1992) and Martínez and Curnow (1992), and became known as the Haley-Knott regression for QTL mapping later. In Haley-Knott regression, the LOD score is calculated as in Equation 3.4, where, instead of maximized likelihood (as in Equation 3.7), residual sum of squares of regression models under different hypothesis are used. Haley-Knott regression provides a good approximation to the standard interval mapping but is much easier in terms of computing. Therefore, it became a very popular method in QTL mapping studies (Broman, 2001; Jansen, 2007).

3.2.2 *i*QTL Mapping

Previous examples used a backcross design to illustrate the basic logic of interval mapping and Haley-Knott regression. In a backcross design, each marker and QTL has only two possible geno-

types, i.e., one of the two possible homozygotes and the heterozygote. According to the definition, analysis of imprinting requires a distinction between maternally and paternally inherited alleles, indicating that back crossing to two parental lines are both needed to assess imprinting. For example, suppose the paternal line is QQ and the maternal line is qq. When backcrossed to the paternal line, the paternally inherited allele in all F_2 individuals will only be Q. Hence, if there is no significant difference between QQ and qQ (maternally inherited allele written first), one may conclude that there would be a maternal imprinting at this locus. However, one cannot confirm a paternal imprinting if a significant difference between these two genotypes is detected since a Mendelian QTL can give the same result. To detect paternal imprinting, therefore, F_1 individuals need to be backcrossed to the maternal line and a paternal imprinting is claimed if no significant difference between QQ and Qq in F_2 individuals is observed. To avoid extra effort on crossing experiment, intercross experiments (e.g., F_2 design) are usually adopted, in which case the two heterozygotes qQ and Qq provide sufficient information for detecting an imprinting effect.

Starting from a Mendelian QTL, the Haley-Knott regression model in an F_2 design is (in matrix form)

$$\boldsymbol{y} = \mathbf{1}m + \boldsymbol{p}_{QQ}\beta_{QQ} + \boldsymbol{p}_{Qq/qQ}\beta_{Qq/qQ} + \boldsymbol{p}_{qq}\beta_{qq} + \boldsymbol{e}, \tag{3.9}$$

where $\boldsymbol{p}$ denotes the vector of probabilities of observing a given genotype, as represented by the subscripts. In a standard quantitative genetic model (e.g., Falconer and Mackay, 1996; Lynch and Walsh, 1998), genotypic values a, d and $-a$ are assigned to genotypes QQ, Qq/qQ and qq. Thus, β_{QQ}, $\beta_{Qq/qQ}$ and β_{qq} in Equation 3.9 represent a, d and $-a$, respectively, and hence it can be rewritten as (Haley and Knott, 1992; Haley *et al.*, 1994)

$$\boldsymbol{y} = \mathbf{1}m + a(\boldsymbol{p}_{QQ} - \boldsymbol{p}_{qq}) + d\boldsymbol{p}_{Qq/qQ} + \boldsymbol{e}. \tag{3.10}$$

Because, when considering parental origins of the inherited allele, $\boldsymbol{p}_{Qq/qQ} = \boldsymbol{p}_{Qq} + \boldsymbol{p}_{qQ}$, Knott *et al.* (1998) wrote Equation 3.10 in a more explicit form, and introduced another term to account for

the imprinting effect i, which initiated iQTL mapping studies. Their model is

$$\boldsymbol{y} = \mathbf{1}m + a(\boldsymbol{p}_{QQ} - \boldsymbol{p}_{qq}) + d(\boldsymbol{p}_{Qq} + \boldsymbol{p}_{qQ}) + i(\boldsymbol{p}_{Qq} - \boldsymbol{p}_{qQ}) + \boldsymbol{e}. \tag{3.11}$$

In this model, the additive effect is defined as the regression coefficient on the contrast between two homozygotes and the imprinting effect is the regression coefficient on the contrast between two heterozygotes, which agrees with the definition of genomic imprinting.

The first applications of Model 3.11 to iQTL mapping were probably studies on muscle-related QTL in pigs (Nezer *et al.*, 1999; Jeon *et al.*, 1999). Both studies detected evidence of paternally expressed QTL at the *IGF2* locus that affected traits like lean meat in ham (%), lean meat mass in ham (kg), lean meat and bone in back (%), weight of internal organs (heart, gram), average backfat depth (mm), and so forth. Specifically, Jeon *et al.* (1999) reported that the QTL explained 30% of the phenotypic variance on lean meat content in ham in the F_2 population. Later, de Koning *et al.* (2000) detected 5 QTL affecting pig body composition traits in a whole-genome scan study using microsatellite markers, 4 of which were deemed to be imprinted: one paternally expressed QTL affecting backfat thickness was mapped to *Sus scrofa* chromosome 2 (SSC2) and a maternally expressed QTL affecting muscle depth was mapped to SSC7; on SSC6, a maternally expressed and a paternally expressed QTL were mapped to the short and long arms, respectively, both affecting intramuscular fat content. Due to a large fraction of the phenotypic variance explained by the individual QTL (up to 10%), the authors claimed that testing for imprinting should become a standard procedure to unravel the genetic control of complex traits and suggested that imprinting in pigs might be more widespread than previously believed.

In de Koning *et al.* (2000), Model 3.11 was reparameterized as

$$\boldsymbol{y} = \mathbf{1}m + a_{\text{♀}}\boldsymbol{p}_{\text{♀}} + a_{\text{♂}}\boldsymbol{p}_{\text{♂}} + d(\boldsymbol{p}_{Qq} + \boldsymbol{p}_{qQ}) + \boldsymbol{e}, \tag{3.12}$$

where $\boldsymbol{p}_{\text{♀}} = \boldsymbol{p}_{QQ} + \boldsymbol{p}_{qQ} - \boldsymbol{p}_{Qq} - \boldsymbol{p}_{qq}$ and $\boldsymbol{p}_{\text{♂}} = \boldsymbol{p}_{QQ} + \boldsymbol{p}_{Qq} - \boldsymbol{p}_{qQ} - \boldsymbol{p}_{qq}$. By doing this, the imprinting

effect i was integrated into two parental additive effects $a_{♀}$ and $a_{♂}$, given the parental origin of the alleles was known. Model 3.12 can be viewed as a prototype of the model of Shete and Amos (2002) (Equation 3.1), and the imprinting pattern (i.e., maternally or paternally) can be inferred by comparing the full model with the following reduced models (de Koning *et al.*, 2000; Rowe *et al.*, 2012)

$$\boldsymbol{y} = \mathbf{1}m + a_{♀}\boldsymbol{p}_{♀} + \boldsymbol{e},$$

$$\boldsymbol{y} = \mathbf{1}m + a_{♂}\boldsymbol{p}_{♂} + \boldsymbol{e}.$$

This procedure rapidly became a basic framework and was adopted in many iQTL mapping studies (de Koning *et al.*, 2001a,b, 2002; Hirooka *et al.*, 2001, 2002). These early applications of iQTL mapping mainly took place in pig crosses, i.e., outbred crosses. Although an impressive achievement was obtained in these studies, Cui *et al.* (2006) argued that, strictly speaking, a QTL identified with outbred crosses is not necessarily an iQTL. This is because if QTL are not fixed in founder lines due to a high degree of heterozygosity, alleles at a given QTL can be different between two outbred parents. As a result, detected differences on genetic effects from paternal and maternal parents may be simply due to different alleles (Lin *et al.*, 2003) rather than an imprinting effect of the same allele. Therefore, Cui *et al.* (2006) proposed an iQTL mapping method under an F_2 design using inbred line crosses. In this method, sex differences in recombination rate during meiosis was used to determine the two heterozygous genotypes. The maximum likelihood method based on a mixture model implemented by the EM algorithm was then employed to detect imprinting effects and imprinting direction in a mouse population. Results suggested that this model provided accurate estimates of the location, direction, effects, and of residual variance for imprinted QTL, and it was later extended to backcross designs (Cui, 2007; Cui *et al.*, 2007) and to functional mapping (Cui *et al.*, 2008) under a similar statistical framework.

iQTL mapping studies were successful for locating genomic regions that are subject to imprinting effect. However, well designed experimental crosses are usually needed for these mapping studies

and QTL detection in a general pedigree is often more challenging. Under this circumstances, a mixed model procedure with the consideration of both a random genetic merit and a fixed QTL effect would be useful. This method started from a general animal model (Henderson, 1975, 1984)

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \qquad \text{(null QTL)} \qquad (3.13)$$

where $\boldsymbol{\beta}$ is a vector of fixed effects with associated incidence matrix $\boldsymbol{X}$; $\boldsymbol{u}$ is a vector of infinitesimal genetic effects with incidence matrix $\boldsymbol{Z}$ and assumed distribution $N(\mathbf{0}, \boldsymbol{A}\sigma_u^2)$, where σ_u^2 is the additive genetic variance and the numerator relationship matrix $\boldsymbol{A}$ links all individuals in the pedigree via kinship; $\boldsymbol{e}$ is the vector of model residual whose elements are assumed independently following a normal distribution with null mean and constant variance σ_e^2. This model contains no genetic effects except for $\boldsymbol{u}$ and hence is referred to as the null QTL model. When QTL are considered, Model 3.13 can be extended in various ways according to different assumptions (Rowe *et al.*, 2012):

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Z}_M q_M + \boldsymbol{e}, \qquad \text{(Mendelian QTL)} \qquad (3.14)$$

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Z}_{\varphi} q_{\varphi} + \boldsymbol{Z}_{\sigma} q_{\sigma} + \boldsymbol{e}, \qquad \text{(full parental QTL)} \qquad (3.15)$$

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Z}_{\varphi} q_{\varphi} + \boldsymbol{e}, \qquad \text{(maternal QTL)} \qquad (3.16)$$

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Z}_{\sigma} q_{\sigma} + \boldsymbol{e}, \qquad \text{(paternal QTL)} \qquad (3.17)$$

where q_M, q_{φ}, q_{σ} are effects of a Mendelian QTL, a maternal QTL, and a paternal QTL with associated incidence matrices $\boldsymbol{Z}_M$, $\boldsymbol{Z}_{\varphi}$, and $\boldsymbol{Z}_{\sigma}$ respectively. Assumptions on the variance of these QTL effects are $Var(q_M) = \boldsymbol{G}_M \sigma_M^2$, $Var(q_{\varphi}) = \boldsymbol{G}_{\varphi} \sigma_{\varphi}^2$, and $Var(q_{\sigma}) = \boldsymbol{G}_{\sigma} \sigma_{\sigma}^2$, where relationship matrices $\boldsymbol{G}_M$, $\boldsymbol{G}_{\varphi}$, and $\boldsymbol{G}_{\sigma}$ for a given QTL position are calculated conditional on flanking marker information and, hence, unique for each position evaluated for a QTL (Liu *et al.*, 2002).

Inference on the existence of QTL of a particular type is then performed by comparing Model 3.13 with one of Models 3.14 – 3.17 using variance components. Further, Model 3.14 can be

viewed as a special case of Model 3.15 with the null hypothesis being that the additive variance is decomposed equally into parts attributed to maternal and paternal parents, and comparing these two models provides a statistic for testing imprinting effects (Hanson *et al.*, 2001; Shete *et al.*, 2003). When separating parental contributions in variance components analysis, the power of *i*QTL detection increased significantly, leading to the conclusion that the variance components method is more powerful than sib pair based methods (Pratt *et al.*, 2000; Hanson *et al.*, 2001). Besides, power was further improved when linkage disequilibrium (LD) information was incorporated into the variance component models (Heuven *et al.*, 2005), as suggested by Lee and van der Werf (2004). Following these early achievements, variance components analysis was successfully applied in QTL detection with the incorporation of parent-of-origin effects in pigs, chicken, and dogs (Heuven *et al.*, 2006; Rowe *et al.*, 2009; Liu *et al.*, 2007). However, it should be noted that population structure had a big impact on *i*QTL mapping such that a spurious uniparental expression would be detected if too few parents were segregating (Heuven *et al.*, 2005; Liu *et al.*, 2007). Therefore, it was suggested that many small families other than fewer larger families should be used to increase the power for detection (Liu *et al.*, 2007). On the other hand, instead of mapping individual *i*QTL, if the focus was on the impact of parent-of-origin effects at an individual's level, variance components approaches have been used in pigs and cattle (de Vries *et al.*, 1994; Kaiser *et al.*, 1998; Engellandt and Tier, 2002; Essl and Voith, 2002), indicating that imprinting was contributing to some complex traits considerably. Specifically, an up to 25% contribution of imprinting to additive variance on many carcass traits has been reported in recent studies (Neugebauer *et al.*, 2010a,b). Because variance component procedures partitioned genetic variation due to imprinting explicitly, in addition to other genetic effects, this approach provides a general framework for analyzing complex traits, as noted recently (Blunk and Reinsch, 2014).

3.3 High-Resolution Whole-Genome Scan for Imprinting Signals

The success of *i*QTL mapping studies improved the understanding of genetic architecture significantly, especially when epigenetic variation is considered in agricultural species. However, similar to conventional QTL mapping studies, a big drawback was that molecular markers used in these *i*QTL mapping studies were usually low-density microsatellite markers, resulting in low-resolution detected regions that potentially contained tens to hundreds of genes. With the advent of high-density SNP (single nucleotide polymorphisms) marker chips at decreasing costs, it became possible to perform whole-genome scans in much larger cohorts at a higher resolution. One advantage of using high-density markers is that with the reduced physical (and hence genetic) distances between adjacent markers, marker genotypes are more informative than in a low-density case, such that designed crosses are less necessary and studies on observed data (e.g., from a human population) can be easily performed (Evans and Cardon, 2004; Schaid *et al.*, 2004). Nowadays, with elegant computational algorithms, it is possible to perform haplotype inference/phasing at a desired accuracy (Browning and Browning, 2011), making whole-genome scan for imprinting signals at the SNP level feasible. Representative studies are Wolf *et al.* (2008a) and Cheverud *et al.* (2008). In these studies, markers showing strong imprinting effects on mouse body weight were detected, and various imprinting patterns based on the results were suggested. Later, using a similar approach in pigs, Coster *et al.* (2012) verified the imprinting status of a litter size related gene *DIO3*, which was in consistent with earlier studies (e.g., da Rocha *et al.*, 2008).

Comparing to a SNP marker, microsatellite markers are usually more polymorphic and hence possessing richer information than a SNP marker, which is biallelic. However, since SNPs are far more numerous than microsatellites over the whole genome, the former is more informative and hence used more often in biological studies. Because the physical distance between SNP markers is much smaller than for microsatellites, a higher resolution mapping can be obtained, explaining why microsatellites have been replaced by SNPs in recent years. Due to this feature and

other characteristics, e.g., decreasing genotyping costs, SNP markers rapidly dominated linkage and association studies, and became the basic tool for genome-wide association studies (GWAS). Theoretically, the whole-genome scan described in Wolf *et al.* (2008a), Cheverud *et al.* (2008) and Coster *et al.* (2012) has the same logic as GWAS, with the major difference being that in GWAS, hypothesis testing is performed over the contrast between two homozygotes and the contrast between two heterozygotes is (usually) not of interest. In the case of imprinting, on the other hand, the latter matters and hence differentiation between Qq and qQ genotypes is necessary. Interestingly, relative to the number of publications on GWAS starting from 2005 (e.g., Hindorff *et al.*, 2014), the number of reports on whole-genome scan for imprinting using SNPs is very low. Possible reasons for this are discussed next.

First, imprinting is still a minor concern in biological and genetic studies. Although review papers on addressing the potential importance of genomic imprinting are available (e.g., Groenen, 2005; Wolf *et al.*, 2008b; Lawson *et al.*, 2013), imprinting attracts less attention than Mendelian genes due to the small number of imprinted genes (Kelsey and Bartolomei, 2012). Second, imprinting is an epigenetic mechanism that does not involve a change of DNA sequence, indicating that the best way of studying imprinting is probably using information from the epigenome level. Because quantification of epigenetic variation became available in 2000's, e.g., methylated DNA immunoprecipitation (MeDIP) and bisulfite sequencing (BS-Seq), many studies used information at the methylation, histon, or RNA levels as study material for imprinting, instead of using SNP polymorphisms. This type of approach includes differential expression studies that use RNA Seq information and epigenome-wide association studies (EWAS) that usually utilize methylation data. As a result, imprinting studies using SNP as input information became less popular. Third, because imprinting displays an unequivalence of paternal and maternal contributions to the offspring, in studies using observational data, information consisting of an affected patient and his/her parents is usually considered, making up a structure called the case-parent trio. This familial information is commonly used in human genetics, by applying a transmission disequilibrium test (TDT) instead

of a whole-genome scan (e.g., Kistner and Weinberg, 2004; Kistner *et al.*, 2006; Hu *et al.*, 2007a,b; Laird and Lange, 2008; Zhou *et al.*, 2015).

The preceding partly explains the shift of paradigm when genomic imprinting is involved and the studies indicated that variation at the epigenome level might be a preferred resource. However, epigenomic profiling is still very expensive at the current time, which may limits its use in large scale studies. Further, most of the genetic variation still resides at the DNA level, suggesting that SNP data should be combined with epigenome profiles as information inputs. In other words, genomic information at the DNA level should not be ignored or abandoned. Therefore, the following chapters start from a whole-genome scale quantitative epigenetic analysis using only SNP information (Chapter 4). Subsequently, in a whole-genome prediction study, SNP serve as the major information input, but the use of epigenetic information is mentioned and discussed (Chapter 5). Finally, methylation data is introduced to perform whole-genome prediction non-parametrically using reproducing kernel Hilbert spaces regression (Chapter 6).

References

Blunk, I. and N. Reinsch, 2014. Genetic variance components when fluctuating imprinting patterns are present. In *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production*. Vancouver, Canada

Broman, K. W., 2001. Review of statistical methods for QTL mapping in experimental crosses. Lab Anim. (NY), 30(7): 44–52

Browning, S. R. and B. L. Browning, 2011. Haplotype phasing: existing methods and new developments. Nat. Rev. Genet., 12(10): 703–714

Cheverud, J. M., R. Hager, C. Roseman, *et al.*, 2008. Genomic imprinting effects on adult body composition in mice. Proc. Natl. Acad. Sci. U.S.A., 105(11): 4253–4258

Coster, A., O. Madsen, H. C. Heuven, *et al.*, 2012. The imprinted gene *DIO3* is a candidate gene for litter size in pigs. PLoS ONE, 7(2): e31825

Cui, Y., 2007. A statistical framework for genome-wide scanning and testing of imprinted quantitative trait loci. J. Theor. Biol., 244(1): 115–126

Cui, Y., J. M. Cheverud, and R. Wu, 2007. A statistical model for dissecting genomic imprinting

through genetic mapping. Genetica, 130(3): 227–239

Cui, Y., S. Li, and G. Li, 2008. Functional mapping imprinted quantitative trait loci underlying developmental characteristics. Theor. Biol. Med. Model, 5: 6

Cui, Y., Q. Lu, J. M. Cheverud, *et al.*, 2006. Model for mapping imprinted quantitative trait loci in an inbred F_2 design. Genomics, 87(4): 543–551

da Rocha, S. T., C. A. Edwards, M. Ito, *et al.*, 2008. Genomic imprinting at the mammalian *Dlk1-Dio3* domain. Trends Genet., 24(6): 306–316

de Koning, D. J., H. Bovenhuis, and J. A. van Arendonk, 2002. On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. Genetics, 161(2): 931–938

de Koning, D. J., B. Harlizius, A. P. Rattink, *et al.*, 2001a. Detection and characterization of quantitative trait loci for meat quality traits in pigs. J. Anim. Sci., 79(11): 2812–2819

de Koning, D. J., A. P. Rattink, B. Harlizius, *et al.*, 2000. Genome-wide scan for body composition in pigs reveals important role of imprinting. Proc. Natl. Acad. Sci. U.S.A., 97(14): 7947–7950

de Koning, D. J., A. P. Rattink, B. Harlizius, *et al.*, 2001b. Detection and characterization of quantitative trait loci for growth and reproduction traits in pigs. Livest. Prod. Sci., 72(3): 185–198

de Vries, A. G., R. Kerr, B. Tier, *et al.*, 1994. Gametic imprinting effects on rate and composition of pig growth. Theor. Appl. Genet., 88(8): 1037–1042

Engellandt, T. H. and B. Tier, 2002. Genetic variances due to imprinted genes in cattle. J. Anim. Breed. Genet., 119(3): 154–165

Essl, A. and K. Voith, 2002. Genomic imprinting effects on dairy- and fitness-related traits in cattle. J. Anim. Breed. Genet., 119(3): 182–189

Evans, D. M. and L. R. Cardon, 2004. Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. Am. J. Hum. Genet., 75(4): 687–692

Falconer, D. S. and T. F. C. Mackay, 1996. *Introduction to Quantitative Genetics.* Prentice Hall, 4th edition

Groenen, M. A. M., 2005. Imprinted QTL in farm animals: a fortuity or a common phenomenon? In L. B. Jorde (Editor), *Encyclopedia of Genetics, Genomics, Proteomics, and Bioinformatics*, volume 1, pp. 304–308. John Wiley & Sons Ltd., Oxford, UK

Haldane, J. B. S., 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. J. Genet., 8: 299–309

Haley, C. S. and S. A. Knott, 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity (Edinb), 69(4): 315–324

Haley, C. S., S. A. Knott, and J. M. Elsen, 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics, 136(3): 1195–1207

Hanson, R. L., S. Kobes, R. S. Lindsay, *et al.*, 2001. Assessment of parent-of-origin effects in linkage analysis of quantitative traits. Am. J. Hum. Genet., 68(4): 951–962

Henderson, C. R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, 31(2): 423–447

Henderson, C. R., 1984. *Applications of Linear Models in Animal Breeding.* University of Guelph, Ontario, Canada

Heuven, H. C. M., H. Bovenhuis, L. L. G. Janss, *et al.*, 2005. Efficiency of population structures for mapping of Mendelian and imprinted quantitative trait loci in outbred pigs using variance component methods. Genet. Sel. Evol., 37(7): 635–655

Heuven, H. C. M., M. J. Eckeveld, N. B. Coster, *et al.*, 2006. Imprinted QTL on *Sus Scrofa* Chromosome 6 affect intramuscular fat percentage in Meishan-F_2 pigs. In *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production.* Belo Horizonte-MG, Brazil

Hindorff, L. A., J. MacArthur, J. Morales, *et al.*, 2014. A catalog of published genome-wide association studies. http://www.genome.gov/gwastudies/

Hirooka, H., D. J. de Koning, B. Harlizius, *et al.*, 2001. A whole-genome scan for quantitative trait loci affecting teat number in pigs. J. Anim. Sci., 79(9): 2320–2326

Hirooka, H., D. J. de Koning, J. A. van Arendonk, *et al.*, 2002. Genome scan reveals new coat color loci in exotic pig cross. J. Hered., 93(1): 1–8

Hu, Y. Q., J. Y. Zhou, and W. K. Fung, 2007a. An extension of the transmission disequilibrium test incorporating imprinting. Genetics, 175(3): 1489–1504

Hu, Y. Q., J. Y. Zhou, F. Sun, *et al.*, 2007b. The transmission disequilibrium test and imprinting effects test based on case-parent pairs. Genet. Epidemiol., 31(4): 273–287

Jansen, R. C., 2007. Chapter 18 - Quantitative trait loci in inbred lines. In D. J. Balding, M. Bishop, and C. Cannings (Editors), *Handbook of Statistical Genetics*, pp. 589–622. John Wiley & Sons, Ltd., 3rd edition

Jeon, J. T., O. Carlborg, A. Tornsten, *et al.*, 1999. A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the *IGF2* locus. Nat. Genet., 21(2): 157–158

Kaiser, C. J., M. E. Goddard, and A. Reverter, 1998. Analysis of gametic imprinting effects for test day milk yield in Australian Holstein cattle. In *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production.* Armidale, Australia

Kelsey, G. and M. S. Bartolomei, 2012. Imprinted genes ... and the number is? PLoS Genet., 8(3): e1002601

Khatib, H., 2007. Is it genomic imprinting or preferential expression? Bioessays, 29(10): 1022–1028

Kistner, E. O., C. Infante-Rivard, and C. R. Weinberg, 2006. A method for using incomplete triads to test maternally mediated genetic effects and parent-of-origin effects in relation to a quantitative trait. Am. J. Epidemiol., 163(3): 255–261

Kistner, E. O. and C. R. Weinberg, 2004. Method for using complete and incomplete trios to identify genes related to a quantitative trait. Genet. Epidemiol., 27(1): 33–42

Knott, S. A., L. Marklund, C. S. Haley, *et al.*, 1998. Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. Genetics, 149(2): 1069–1080

Laird, N. M. and C. Lange, 2008. Family-based methods for linkage and association analysis. Adv. Genet., 60: 219–252

Lander, E. S. and D. Botstein, 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121(1): 185–199

Lawson, H. A., J. M. Cheverud, and J. B. Wolf, 2013. Genomic imprinting and parent-of-origin effects on complex traits. Nat. Rev. Genet., 14(9): 609–617

Lee, S. H. and J. H. van der Werf, 2004. The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage. Genet. Sel. Evol., 36(2): 145–161

Lin, M., X. Y. Lou, M. Chang, *et al.*, 2003. A general statistical framework for mapping quantitative trait loci in nonmodel systems: issue for characterizing linkage phases. Genetics, 165(7): 901–913

Liu, T., R. J. Todhunter, S. Wu, *et al.*, 2007. A random model for mapping imprinted quantitative trait loci in a structured pedigree: an implication for mapping canine hip dysplasia. Genomics, 90(2): 276–284

Liu, Y., G. B. Jansen, and C. Y. Lin, 2002. The covariance between relatives conditional on genetic markers. Genet. Sel. Evol., 34(6): 657–678

Lynch, M. and B. Walsh, 1998. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates

Martínez, O. and R. N. Curnow, 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet., 85(4): 480–488

Morcos, L., B. Ge, V. Koka, *et al.*, 2011. Genome-wide assessment of imprinted expression in human cells. Genome Biol., 12(3): R25

Neugebauer, N., H. Luther, and N. Reinsch, 2010a. Parent-of-origin effects cause genetic variation in pig performance traits. Animal, 4(5): 672–681

Neugebauer, N., I. Räder, H. J. Schild, *et al.*, 2010b. Evidence for parent-of-origin effects on genetic variability of beef traits. J. Anim. Sci., 88(2): 523–532

Nezer, C., L. Moreau, B. Brouwers, *et al.*, 1999. An imprinted QTL with major effect on muscle mass and fat deposition maps to the *IGF2* locus in pigs. Nat. Genet., 21(2): 155–156

Pratt, S. C., M. J. Daly, and L. Kruglyak, 2000. Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. Am. J. Hum. Genet., 66(3): 1153–1157

Rowe, S., S. Bishop, and D. J. de Koning, 2012. Imprinting in genome analysis: modeling parent-of-origin effects in QTL studies. In H. Khatib (Editor), *Livestock Epigenetics*, pp. 113–129. John Wiley & Sons

Rowe, S. J., R. Pong-Wong, C. S. Haley, *et al.*, 2009. Detecting parent of origin and dominant QTL in a two-generation commercial poultry pedigree using variance component methodology. Genet. Sel. Evol., 41(1): 6

Schaid, D. J., J. C. Guenther, G. B. Christensen, *et al.*, 2004. Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci. Am. J. Hum. Genet., 75(6): 948–965

Shete, S. and C. I. Amos, 2002. Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. Am. J. Hum. Genet., 70(3): 751–757

Shete, S., X. Zhou, and g. p. s. I. Amos, 2003. Genomic imprinting and linkage test for quantitative-trait loci in extended pedigrees. Am. J. Hum. Genet., 73(4): 933–938

Soller, M., T. Brody, and A. Genizi, 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet., 47(1): 35–39

Spencer, H. G., 2002. The correlation between relatives on the supposition of genomic imprinting. Genetics, 161(1): 411–417

Wolf, J. B., J. M. Cheverud, C. Roseman, *et al.*, 2008a. Genome-wide analysis reveals a complex pattern of genomic imprinting in mice. PLoS Genet., 4(6): e1000091

Wolf, J. B., R. Hager, and J. M. Cheverud, 2008b. Genomic imprinting effects on complex traits: a phenotype-based perspective. Epigenetics, 3(6): 295–299

Zhou, J. Y., H. Q. He, X. P. You, *et al.*, 2015. A powerful association test for qualitative traits incorporating imprinting effects using general pedigree data. J. Hum. Genet., 60(2): 77–83

Chapter 4

Impact of Imprinting on the Genetic Variation of Mouse Body Mass Index

In the previous chapter, *i*QTL mapping and whole-genome scan studies with consideration of imprinting were introduced, which were mainly motivated by the potential impact of imprinting on complex traits. However, little is known about the consequences when existing imprinting is ignored in the genetic evaluation process. In this chapter, a GWAS-like scan was performed to address this question. Results suggested that ignoring existing imprinting may underestimate the proportion of transmissible genetic variation, which may partly explain the missing heritability problem from an epigenetic perspective.

4.1 Introduction

Genomic imprinting, an epigenetic process, is the preferential or differential gene expression in a parent-of-origin fashion, as described in Chapter 2. The evolution and precise mechanisms of genomic imprinting are not fully understood yet, but it is believed that imprinting can affect many complex traits (Kilpinen and Dermitzakis, 2012), including economically important traits

in agricultural animal and plant species (Georges *et al.*, 2003; Wolf *et al.*, 2008b; Spencer, 2009; Lawson *et al.*, 2013) as well as human diseases like the Prader-Willi (PWS) and Angelman (AS) syndromes (Meijers-Heijboer *et al.*, 1992; Nicholls *et al.*, 1998). Thus, relating phenotype and its variation of complex traits to (epi)genetic variants is of interest.

Studies of complex traits often aim at partitioning phenotypic variance into different components. In classical quantitative genetics (e.g., Falconer and Mackay, 1996), the phenotypic variance is partitioned into the sum of genotypic and environmental variances, and the genotypic variance is further subdivided into additive and dominance variances. If two or more loci are involved, there could also be an epistatic variance component. The ratio between the additive genetic and the phenotypic variances (σ^2_A and σ^2_P, respectively) defines the narrow sense heritability (h^2). Therefore, h^2 is usually interpreted as the proportion of phenotypic variance explained by additive variance. This parameter is also the expected fraction of the selection differential transmitted from the parental to the offspring generation, and is crucial in artificial selection since it determines the mean performance of the offspring generation after applying selection to the parental generation (Falconer and Mackay, 1996). Hence, knowledge of h^2 is important in genetic improvement programs for predicting breeding values of selection candidates.

Conventionally, h^2 can be estimated using phenotypic records and kinship information with likelihood-based or Bayesian methods under a linear mixed effect model specification (Fisher, 1918; Kempthorne, 1954; Searle, 1971; Henderson, 1984; Searle *et al.*, 2006; Sorensen and Gianola, 2002). In recent years, the advent of SNP (single nucleotide polymorphisms) markers made it possible to perform genetic analysis at the DNA level as well as to carry out genome-wide association studies (GWAS), with the goal being finding genomic regions that potentially have an effect on a complex trait of interest. Under a GWAS framework, one can estimate the substitution effect of an allele at some known locus and use an estimate of $2pq\alpha^2$ as the additive variance contributed by that locus if Hardy-Weinberg equilibrium holds, and, therefore, an estimate of "marker-based" heritability can be obtained. Usually, additive variation attracted most attention with dominance in the background

scene because it is deemed not to contribute to heritable variation under a classical quantitative genetics framework (e.g., Falconer and Mackay, 1996; Lynch and Walsh, 1998). However, unlike dominance or epistasis involving dominance, an imprinting effect is thought to be transmissible over generations (Reik and Walter, 2001; Spencer, 2002). According to the one-locus imprinting model introduced in Section 3.1, additive genetic variation increases in the presence of genomic imprinting. Therefore, imprinting adds a transmissible variation to the genetic variance and it contributes to narrow sense heritability by definition. In what follows, a stylized analysis is provided to describe the impact of imprinting on genetic variation under various circumstances. Then a GWAS-like whole-genome scan is performed on a mouse data to address several topics related to genetic variation under imprinting.

4.2 Contribution of Imprinting to Genetic Variation

The potential role of epigenetics on complex traits has led to studies of the impact of epigenetic variation on phenotypic and genetic variations. In a recent *in silico* study, for example, it was shown that epigenetic modification of one allele at a biallelic locus can result in an 11% of total genetic variance attributed to epigenetic variation at moderate allele frequency even if u, the epigenetic modification rate, is as low as 0.01 (Wang *et al.*, 2012). The proportion of genetic variance explained by epigenetic variation could be as large as 18% if u increases to 0.5. This can be explained by viewing the epigenetic modification as producing an epi-mutation that has a similar effect as a regular mutation event if the epi-mutation persists a relatively long time in a population, i.e., if transmissible between generations.

When genomic imprinting is discussed here as a specific epigenetic phenomenon, recall from Chapter 3 that the genetic variance at a biallelic imprinted locus is

$$\sigma_G^2 = 2pq(a + d(q - p))^2 + (2pqd)^2 + 2pqi^2 \tag{4.1}$$

and the additive genetic variance is

$$\sigma_A^2 = \sigma_{♀}^2 + \sigma_{♂}^2 = pq\alpha_{♀}^2 + pq\alpha_{♂}^2 = 2pq(a + d(q-p))^2 + 2pqi^2. \tag{4.2}$$

Hence, additive genetic variance increases by $2pqi^2$ when imprinting is present. For the sake of clarify, $2pq(a+d(q-p))^2+2pqi^2$ is termed as the additive genetic variance in subsequent discussion, because this is the variance between breeding values under imprinting. Parts $2pq(a + d(q - p))^2$ and $2pqi^2$ are referred to as Mendelian (i.e., the unimprinted part) and imprinting variances, and are denoted by σ_{Men}^2 and σ_{Imp}^2, respectively (Neugebauer *et al.*, 2010). Analogous to the definition of h^2, the ratio $\sigma_{Imp}^2/\sigma_G^2$ defines the proportion of total genetic variance explained by imprinting. This ratio is, to some extent, equivalent to the definition of R_e^2 in Wang *et al.* (2012), with the only difference being that these authors were interested in a broader concept of epigenetic mechanism while here we are interested in imprinting only. We graphically illustrate how imprinting can impact the evaluation of marked variance, and its consequences if ignored. We set $a = 2$ and let the imprinting effect i vary between 0 (no imprinting) and a (complete imprinting) according to the previously described imprinting model. Four different values were assigned to the dominance effect d: 0, $\frac{1}{4}a$, $\frac{1}{2}a$ and a, representing from no dominance to complete dominance. Allele frequency p for the A_1 allele varied from 0 to 1.

As shown in Figure 4.1, $R_e^2 = \sigma_{Imp}^2/\sigma_G^2$ increases with i, d and p. When $d = 0$, R_e^2 does not vary with p since in this case

$$R_e^2 = \frac{2pqi^2}{2pqa^2 + 2pqi^2} = \frac{i^2}{a^2 + i^2}.$$

Under dominance, allele frequency drives R_e^2 from small values at lower allele frequency to large values at higher frequency, with more pronounced effects with larger values of d. When imprinting effects are small (e.g., $i < \frac{1}{4}a$), it seldom accounts for more than 10% of the genetic variance, unless p is close to 1 and d is close to a.

Figure 4.2 shows how narrow sense heritability changes with (imprinting model) or without

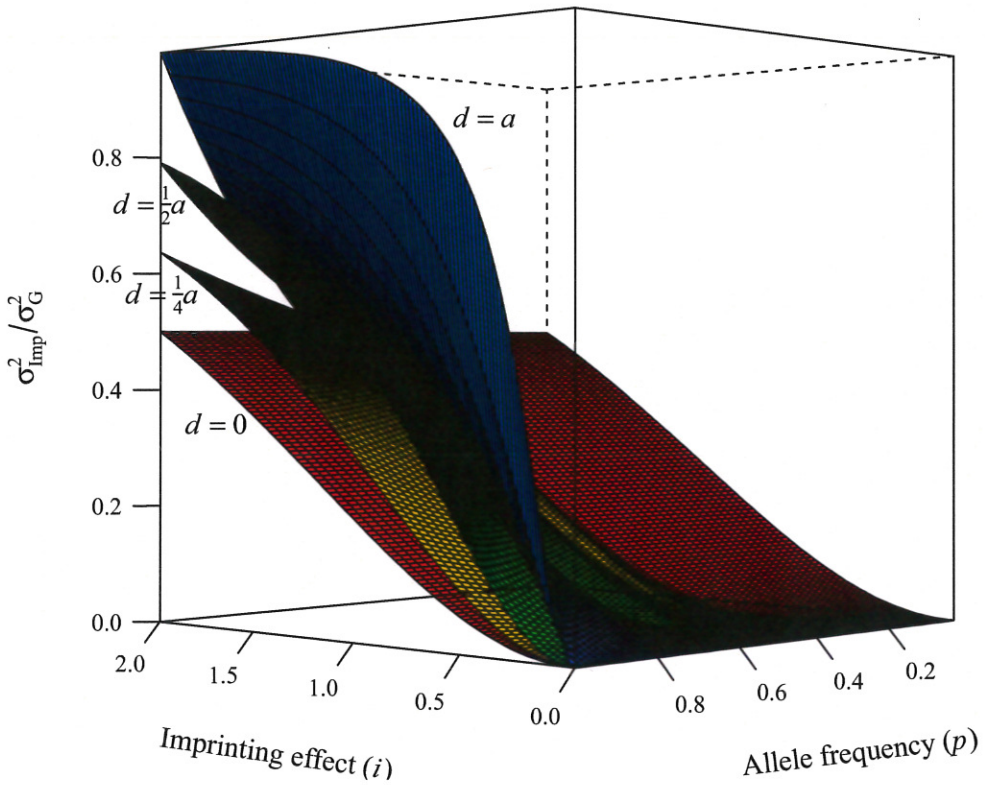


Figure 4.1: Proportion of genetic variance contributed by imprinting as a function of allele frequency (p), dominance level (d), and imprinting effect (i).

(additive model) consideration of imprinting at the four values of d. The environmental variance was set to $\sigma_e^2 = 4$ across all situations and it was assumed that there was no interaction between environmental and genetic factors; a, i, d and p were as before. The additive variance obtained with imprinting was always larger than when an additive model was employed, as expected by construction. If we denote "epigenetic heritability" as h_e^2 (Wang *et al.*, 2012) and that without consideration of imprinting as h^2, the difference between h_e^2 and h^2 is maximum when imprinting is at its highest level. This is not surprising because the larger i is, the higher the proportion of additive variation (i.e., σ_A^2) accounted for by imprinting is (Equation 4.2). Thus, if imprinting is present, the standard additive model would capture only part of the additive variance, resulting in an underestimate of the potentially markable variation.

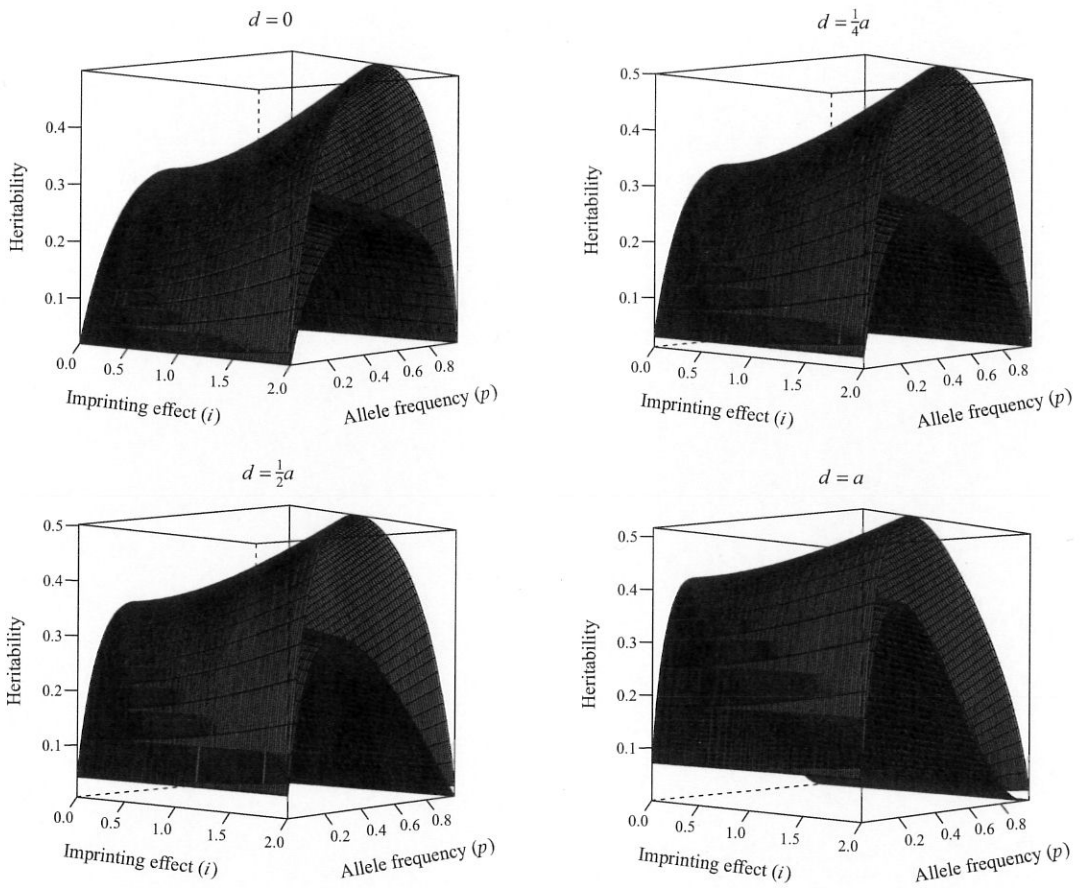


Figure 4.2: Narrow sense heritability with (h_e^2, red surface) or without (h^2, blue surface) consideration of imprinting as a function of allele frequency (p) and imprinting effect (i) at various level of dominance.

4.3 Mouse Data Analysis – Materials and Methods

The preceding discussion on the quantitative imprinting model applies to a single locus only and does not guide on how imprinting would contribute to heritability of a complex trait, presumably affected by multiple loci and with many of these not imprinted at all. Also, it is unknown how imprinting affects estimates of heritability when it is ignored in the estimation procedure. Here, a real data analysis was performed to evaluate the impact of imprinting on a quantitative trait.

Previous studies have suggested that obesity-related traits could be affected by imprinting in

both humans (Gorlova *et al.*, 2003; Dong *et al.*, 2005) and mice (Rance *et al.*, 2005). Hence, mouse body mass index (BMI) was chosen as the target trait in this analysis, considered to be a good indicator of obesity status, as response variable in this analysis. The mouse dataset (build 37), generated for a series of studies on obesity and diabetes, was downloaded from The Wellcome Trust Centre for Human Genetics website (`http://mus.well.ox.ac.uk/mouse/HS/`). This population was obtained by crossing eight inbred strains followed by 50 generations of approximately random mating. BMI measurements were pre-corrected for body weight, season, month and day for a total of 1,940 F_2 individuals (168 full-sib families), with more than 12,000 genotyped SNP markers located on 19 autosomes. BMI values seemed normally distributed with mean -0.4568 (negative values were due to pre-correction on original data) and variance 0.0357. Additional descriptions of the data are in the website and in Valdar *et al.* (2006).

To test the effect of imprinting, one must be able to distinguish two heterozygous genotypes, which is impossible if conventional coding systems used in GWAS or whole-genome prediction studies (e.g., genotypes A_2A_2, A_1A_2/A_2A_1 and A_1A_1 coded as -1, 0, and 1, respectively) are adopted, because A_1A_2 and A_2A_1 are not differentiated. To make A_1A_2 and A_2A_1 distinguishable, marker genotypes (in the form of AA, AB, BB) were fed into BEAGLE 3.3.2 (Browning and Browning, 2009; Browning, 2011) for sporadic missing genotype imputation and haplotype phase inference. This software can perform haplotype inference of unphased (unknown parental origin) genotypic data using linkage information between marker genotypes, with or without pedigree information, and give an inferred phased (known parental origin status) genotype as an output. With phased genotype, markers can be coded as described below. After filtering markers with minor allele frequency (MAF) less than 0.05, 10,021 markers were kept for analysis.

One objective of this study is to assess the consequences of "erroneously" using an additive model without considering imprinting in GWAS if imprinting does affect that trait. Therefore, the data was analyzed using regression models with or without imprinting, as described below. First, according to the imprinting model introduced in Section 3.1, the following matrix can be used to

associate different genetic effects with the four possible genotypes (Wolf *et al.*, 2008a; Cheverud *et al.*, 2008; Coster *et al.*, 2012):

$$\boldsymbol{S} = \begin{array}{c} \begin{matrix} \boldsymbol{I}_a & \boldsymbol{I}_d & \boldsymbol{I}_i \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \\ -1 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} \longleftarrow QQ \\ \longleftarrow qQ \\ \longleftarrow Qq \\ \longleftarrow qq \end{matrix}, \tag{4.3}$$

where $\boldsymbol{I}_a$, $\boldsymbol{I}_d$, and $\boldsymbol{I}_i$ are vector indicators for the additive (a), dominance (d), and imprinting (i) effects in the four genotypes, respectively. Using this coding matrix, models with an additive effect only, additive and dominance, and additive plus dominance plus imprinting can be written in matrix form as:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{I}_a\beta_1 \qquad + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Q}\boldsymbol{c} + \boldsymbol{e}, \tag{4.4}$$

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{I}_a\beta_1 + \boldsymbol{I}_d\beta_2 \qquad + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Q}\boldsymbol{c} + \boldsymbol{e}, \tag{4.5}$$

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{I}_a\beta_1 + \boldsymbol{I}_d\beta_2 + \boldsymbol{I}_i\beta_3 + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Q}\boldsymbol{c} + \boldsymbol{e}, \tag{4.6}$$

where $\boldsymbol{y}$ is an n-element vector containing the observations; μ is the population mean common to all individuals; $\boldsymbol{X}$ is the incidence matrix relating the vector $\boldsymbol{y}$ to the vector of fixed effects $\boldsymbol{b}$ (sex, litter size and cage density); β_1, β_2 and β_3 are regression coefficients that are interpreted as additive, dominance, and imprinting effects, respectively; $\boldsymbol{u}$ is the vector of infinitesimal additive effect with associated incidence matrix $\boldsymbol{Z}$, and it is assumed that $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{A}\sigma_u^2)$, with $\boldsymbol{A}$ being the additive relationship matrix constructed from the pedigree and σ_u^2 the infinitesimal additive genetic variance; $\boldsymbol{c}$, with associated incidence matrix $\boldsymbol{Q}$, is the vector of normally and independently distributed random effects represented by different cages in which an individual is raised, and it is assumed that $\boldsymbol{c}$ has a zero mean and homogeneous variance σ_c^2; $\boldsymbol{e}$ is the vector of model residuals, whose elements are assumed to be normally and independently distributed with zero mean and

homogeneous variance σ_e^2.

A likelihood ratio test (LRT) between Models 4.6 and 4.5 tests significance of β_3, which represents the imprinting effect i; a LRT between Models 4.5 and 4.4 tests significance of β_2, the dominance effect d; and a LRT between Model 4.4 and a null model without marker information tests significance of β_1, interpreted here as the allelic substitution effect α. This procedure of data analysis is graphically represented in Figure 4.3. The main objective of this study was to compare a model with imprinting with the common GWAS strategy used today (i.e., considering additive but not dominance effect) to evaluate the extent to which imprinting affects inference on marked variance. The marked variance ignoring imprinting was assessed as

$$\hat{\sigma}^2_{\text{SNP}} = \sum_{j \in \text{Box 2}} \hat{\sigma}^2_{Men,j} = \sum_{j \in \text{Box 2}} 2\hat{p}_j\hat{q}_j\hat{\alpha}^2_j, \tag{4.7}$$

using only markers falling in Box 2 of Figure 4.3, where $\hat{p}_j$ and $\hat{q}_j = 1 - \hat{p}_j$ are maximum likelihood estimates of allelic frequencies at marker locus j. With consideration of imprinting, the marked variance would be

$$\hat{\sigma}^2_{\text{SNP}} = \sum_{j \in \text{Box 2}} \hat{\sigma}^2_{Men,j} + \sum_{j \in \text{Box 1}} \left(\hat{\sigma}^2_{Men,j} + \hat{\sigma}^2_{Imp,j}\right) = \sum_{j \in \text{Box 2}} 2\hat{p}_j\hat{q}_j\hat{\alpha}^2_j + \sum_{j \in \text{Box 1}} 2\hat{p}_j\hat{q}_j\left(\hat{\alpha}^2_j + \hat{i}^2_j\right), \tag{4.8}$$

using "imprinted markers" (Box 1) and "unimprinted markers" (Box 2). In both cases, linkage equilibrium between markers was assumed.

In order to deal with potential problems raised by multiple testing in single marker regression, the p-value for individual testing was set to 1.316×10^{-5} to obtain a 0.05 genome-wide type I error rate using the Šidák's correction. The effective number of independent tests used in the multiple testing correction was calculated using LD information between markers based on the method described in Moskvina and Schmidt (2008). LD (measured by r, the pairwise haplotypic Pearson's correlation coefficient) between marker pairs across the whole genome was calculated using the R package genetics (Warnes *et al.*, 2012). Models were fitted using R package pedigreemm (Vázquez

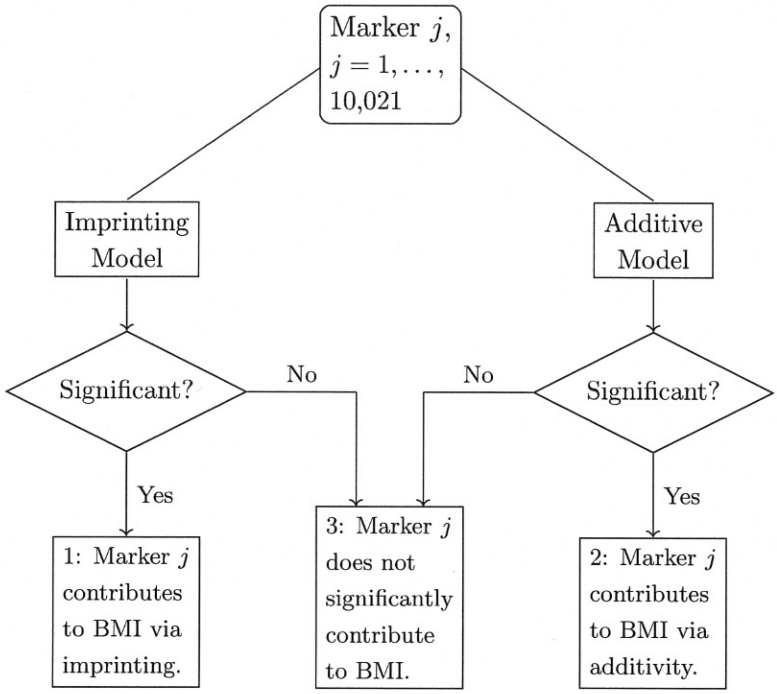


Figure 4.3: Workflow for data analysis.

et al., 2010; Bates and Vázquez, 2013) with variance components of random effects estimated via restricted maximum likelihood (REML).

4.4 Results and Discussion

4.4.1 Significant Markers and Marked Variance

After data cleaning, 10,021 SNP markers were kept for the whole genome scan using methods described in the previous section. As a result, 7 markers were additively significant, and 11 markers were significant for an imprinting effect, either from the paternal side or from the maternal side. The latter suggests the markers are linked with imprinted genes or QTLs. Therefore, all discussions of "imprinted markers" hereafter should be interpreted accordingly. Because many adjacent markers showed co-significance due to high LD (r^2 between markers > 0.99), information redundancy exists for these markers. In order to assess the variance explained by each genomic region, we chose only

one marker from each highly correlated marker cluster. After this filtering, only 3 markers were additively significant and 6 were significant for imprinting, 4 of which were paternally imprinted and 2 were maternally imprinted (Table 4.1). Imprinting direction (i.e., maternal imprinting or paternal imprinting) was determined from the signs of $\hat{\alpha}$ and $\hat{i}$. This is because, according to the imprinting model and the genotype codes, the maternal and paternal allelic substitution effects can be written as $\alpha - i$ and $\alpha + i$. Since reduced expression induced by imprinting indicates a smaller absolute value of the parental substitution effect, a maternal imprinting is then suggested if $\hat{\alpha}$ and $\hat{i}$ have the same sign, and a paternal imprinting is suggested if these two estimates have different signs. This depends on how the four genotypes are coded and one may obtain a reversed result if imprinting is coded oppositely. With these "uniquely" significant markers, marked variance was then computed according to Equations 4.7 and 4.8. We found that $\hat{\sigma}^2_{\text{SNP}}$ with and without consideration of imprinting was $1.218 \times 10^{-4} + 0.624 \times 10^{-4} = 1.842 \times 10^{-4}$ and 1.218×10^{-4}, respectively. The variances explained by the infinitesimal and random cage effects were 3.816×10^{-4} and 4.742×10^{-4} for the imprinting model, and 3.805×10^{-4} and 4.747×10^{-4} for the additive model, respectively. Residual variance was about 1.869×10^{-3} for both cases. Since there were estimates of variance components (i.e., $\hat{\sigma}^2_u$, $\hat{\sigma}^2_c$ and $\hat{\sigma}^2_e$) for every marker, the estimates reported here were the average over 10,021 estimates. Small standard deviations of REML estimates in both models indicated that estimates were fairly precise. Variance components estimates are presented in Table 4.2 along with their asymptotic standard deviations. Values in Table 4.2 indicated that marked variance was increased by 50% in this GWAS-like whole genome scan if variation due to imprinting was considered. In other words, if existing imprinting was not accounted for, about one third marked variation would be lost, potentially leading to wrong conclusions in genetic analysis using SNP markers.

Table 4.1: Significant markers and imprinting status when imprinting was accounted for.

Marker	Chr.	Closest QTL[1]	Status[2]	p-value[3]	$\hat{\alpha}$ ($\times 10^{-3}$)	$\hat{i}$ ($\times 10^{-3}$)
rs3697020	2	*T2dm3* (within)	A	1.476×10^{-6}	11.63 ± 2.41	-
rs3676388	2	*T2dm2sa* (within)	A	9.357×10^{-6}	-9.99 ± 2.26	-
rs3726626	15	*W3q6* (within)	A	2.987×10^{-6}	9.30 ± 1.97	-
rs3662117	2	*Gnf1* (within)	M	4.061×10^{-6}	2.51 ± 2.61	6.20 ± 2.10
rs6212614	3	*Orgwq5* (within)	M	6.197×10^{-6}	2.11 ± 1.96	4.18 ± 1.59
rs13476734	2	*T2dm3* (within)	P	1.302×10^{-5}	-1.36 ± 2.05	4.59 ± 1.79
rs6371982	3	*W10q3* (within)	P	5.055×10^{-6}	1.05 ± 1.94	-4.36 ± 1.59
rs3665109	3	*W10q3* (within)	P	8.690×10^{-6}	-1.82 ± 1.92	4.00 ± 1.62
gnf04.110.360	4	*W10q10* (within)	P	9.365×10^{-6}	1.79 ± 1.96	-4.17 ± 1.62

[1] Information from the Mouse Genome Informatics website (`http://www.informatics.jax.org/`).
[2] A: additively significant; M: maternal imprinting; P: paternal imprinting.
[3] p-value threshold was set to 1.316×10^{-5} to ensure a 0.05 whole-genome type I error rate with Šidák's correction.

Table 4.2: Variance components estimates ($\times 10^{-4}$) using models with (Imp) or without (Add) imprinting effect.

Model	$\hat{\sigma}^2_{\text{SNP}}$	$\hat{\sigma}^2_u$ (infinitesimal)	$\hat{\sigma}^2_c$ (cage)	$\hat{\sigma}^2_e$ (residual)
Imp	1.842	3.816 ± 0.762	4.742 ± 0.344	18.687 ± 1.36
Add	1.218	3.805 ± 0.793	4.747 ± 0.360	18.694 ± 1.37

4.4.2 Interpretation of Significant Markers

In this study, we found 3 markers that are additively related to mouse BMI, all of which are related to a certain QTL that has an effect on mouse body weight or diabetes. Particularly, marker rs3697020 is located in a diabetes related QTL *T2dm3* (type 2 diabetes mellitus 3, chromosome 2) that is also highly interactive with obesity (Stoehr *et al.*, 2000); marker rs3676388 is located in another diabetes related QTL *T2dm2sa* (type 2 diabetes mellitus 2 in SMXA RI mice) on the same chromosome (Kobayashi *et al.*, 2006); lastly, marker rs3726626 is located in a body weight related QTL *W3q6* (weight 3 weeks QTL 6) on chromosome 15 (Rocha *et al.*, 2004). Although the main effect of QTL *T2dm3* and *T2dm2sa* is related to the development of type II diabetes in mice, both are highly correlated with obesity status in mice (Stoehr *et al.*, 2000; Kobayashi *et al.*, 2006), which is commonly considered as a high risk of developing diabetes. Since the data used here was generated for a series of studies on mice diabetes, it was not surprising that markers associated

with diabetes-related QTL were detected.

All 6 presumably imprinted markers detected in our analysis are located in the vicinity of QTLs associated with body weight or growth. For example, marker rs3662117 is in *Gnf1* (growth and fatness 1), a QTL located on chromosome 2 that has a large impact on growth and body composition (Jerez-Timaure *et al.*, 2005). Marker rs6212614 resides in *Orgwq5* (organ weight QTL 5, chromosome 3), a QTL affecting organ weight in mouse (Leamy *et al.*, 2002). This pleiotropic QTL has an impact on limb bone length as well, such that it may potentially affect body length and hence influence body mass index. Markers rs6371982, rs3665109, and gnf04.110.360 are located in *W10q3* (weight 10 weeks QTL 3) and *W10q10* (weight 10 weeks QTL 10) on chromosomes 3 and 4 respectively, which are two QTLs affecting mouse body weight at the age of 10 weeks (Rocha *et al.*, 2004). Interestingly, markers rs3697020 and rs13476734 are both in QTL *T2dm3*, but one is additively significant and the other has a strong imprinting effect. Since the distance between these two markers is large (about 5 Mb), it is possible that these two markers are capturing different signals (see below). Same as the additive markers, locations of these presumably imprinted markers indicated that variation on BMI is likely an inherited feature of variation on body weight and body length via the major QTLs.

We also checked whether these 6 presumably imprinted markers are located in any known imprinted regions. It was found that 5 out of 6 are in the genomic region of reported imprinted genes or *i*QTLs (imprinted QTL). Specifically, markers rs6371982 (chromosome 3, 16.96 cM) and rs3665109 (chromosome 3, 19.81 cM) are both in the range of *i*QTL *Wti3.1* (chromosome 3, 3.79~32.75 cM), which has a significant effect on most mouse body weights from week 1 to 9 and is expressed from the maternally inherited allele (Wolf *et al.*, 2008a). Marker rs6212614 (chromosome 3, 60.92 cM) is located in the range of another weight related *i*QTL *Wti3.2* (chromoeome 3, 60.71 cM), which was also reported in Wolf *et al.* (2008a). Marker rs13476734 (chromosome 2, 60.01 cM) is adjacent to a maternally expressed imprinted gene *Gatm* (glycine amidinotransferase, 60.63 cM) (Williamson *et al.*, 2013). This gene encodes a metabolic enzyme involved in creatine synthesis,

which plays an important role in embryonic and fetal growth as well as brain functioning (Sandell *et al.*, 2003). Marker rs3662117 (chromosome 2, 75.95 cM) is adjacent to a paternally expressed protein coding gene *Mcts2* (malignant T cell amplified sequence 2, 75.41 cM) (Williamson *et al.*, 2013), which influences the choice of polyadenylation (poly A) site for transcripts of the host gene *H13* in an allele-specific manner (Wood *et al.*, 2008); but no strong evidence regarding the effect of *Mcts2* on body weight, obesity, or diabetes has been reported. Besides these five markers, marker gnf04.110.360 (chromosome 4, 56.49 cM) does not fall in any known imprinted region. However, it is located in a genomic region that is predicted to harbor three maternally expressed genes (Luedi *et al.*, 2005). These genes are *4931406I20Rik* (53.44 cM), *Krc* (55.51 cM), and *Grik3* (58.91 cM). There are also two genes adjacent to this interval that are predicted to be paternally expressed (*Ftl2* and *AU040320*), but the positions of these two genes are outside of the maternally expressed region (58.94 cM and 60.94 cM, respectively). Therefore, these two intervals are likely two adjacent clusters that have different imprinting directions, and the imprinting direction of this marker detected in our study matched with previous findings. Unfortunately, no evidence indicating association between body weight and these three genes has been reported.

Our analysis indicated that in this particular data set, markers associated with mouse BMI through either additivity or imprinting can be effectively detected, and the functions of the genes or QTLs harboring these markers supported our discovery on the marker-trait association. Elevated estimates of marked variances suggested that, by incorporating imprinting effects in to a quantitative genetic model, the proportion of phenotypic variance explained by significant markers increased noticeably. In addition to three markers detected using the additive model, six markers were deemed associated with an imprinting effect when this phenomenon was accounted for; the directions of the imprinting effects of all six markers were consistent with previously reported studies. This indicated that the imprinting model detected extra variation that the additive model was not able to capture, so a higher estimate of marked variance was obtained. However, this result was achieved by adding variances contributed by markers from distinct single marker regression models,

which may give a misleading picture of the variance captured by markers because LD between them may overemphasize the contribution of significant markers (Gianola *et al.*, 2013). Although only one marker in each high LD cluster was kept for calculating marked variance in order to reduce bias, caution still needs to be exercised when interpreting this variance since it was obtained from unshrunken estimates of marker effects with simple regression approaches.

4.4.3 Validation of Imprinting Detection – A Simulation

Besides a potential inflation of marker-explained variance stemming from LD between markers, it should also be noted that the detection of imprinting relies mainly on the comparison between heterozygotes, which might be confounded by dominance under some circumstances. For example, even though columns $\boldsymbol{I}_d$ and $\boldsymbol{I}_i$ in the $\boldsymbol{S}$ matrix (Equation 4.3) are ideally orthogonal, there might be a large collinearity if heterozygotes are mostly say, A_1A_2, and hence hampering estimability of either the dominance or the imprinting effect. If, on the other hand, the two heterozygote types have similar frequencies in the population, both dominance and imprinting effects are identifiable and the estimates of the two effects would be uncorrelated, ideally. Thus, the results presented in the previous section would be more convincing if the detection of imprinting was not affected by dominance.

In order to test for potential confounding between imprinting and dominance, we performed the following simulation. First, a population of 5,000 unrelated individuals was created. For each individual, we generated 500 biallelic loci in linkage equilibrium, with allele frequencies varying over $\{0.05, 0.10, 0.15, \ldots, 0.90, 0.95\}$. One hundred out of the 500 loci were randomly selected to have additive effects generated from a standard normal distribution. Within these 100 loci, 50 and 10 were randomly chosen to have dominance and imprinting effects, respectively, both generated from a standard normal distribution. Note that some loci may have all three true effects since we did not force the two sets of loci with either dominance or imprinting effects to be mutually exclusive. Genotypic values at each of the loci that had an effect were created according to Figure 3.1, given

the genotype at that locus. Environmental effects were drawn from a normal distribution with zero mean and variance equal to the variance among genotypic values so the heritability was roughly 0.5.

We fitted Models 4.5 and 4.6 to the simulated data, as well as the following model where the dominance effect was not accounted for

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{I}_a\beta_1 + \boldsymbol{I}_i\beta_3 + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Q}\boldsymbol{c} + \boldsymbol{e}. \tag{4.9}$$

The reason for fitting Model 4.9 is that, because $\boldsymbol{I}_d$ and $\boldsymbol{I}_i$ are orthogonal, we expect the estimate of β_3 (representing i) from this model should be equal to that obtained from Model 4.6, conditionally on the additive effect β_1. We then compared the estimates of imprinting effects from Models 4.6 and 4.9 and dominance effects from Models 4.6 and 4.5. As shown in Figure 4.4, regardless of whether estimated separately or jointly, the dominance and imprinting effects were uncorrelated to each other, reflecting that the population is under Hardy-Weinberg equilibrium. When using the real mouse data, the same picture emerged (Figure 4.5). Therefore, inferences on the imprinting effect in this current data set are unlikely to be confounded by dominance.

Apart from a potential confounding between imprinting and dominance, we were also interested in testing whether the LRT that was applied to test for significant imprinting effect would pick up any unexisting imprinting effect as a false discovery. To do this, we took the same simulated population as described above but generated the genotypic values by including only simulated additive and dominance effects (i.e., without adding the simulated imprinting effect). We denote this data as the Dom data and referred to the one with true imprinting effects as the Imp data. Then we fitted Models 4.4 and 4.9 to the Dom data to evaluate how imprinting could be detected in a population not affected by imprinting and compared to the result obtained when the same procedure was applied on the Imp data. As a result, one locus showing significant imprinting effect was detected in the Imp data (p-value 3.49×10^{-16}) and none were detected using the Dom data,

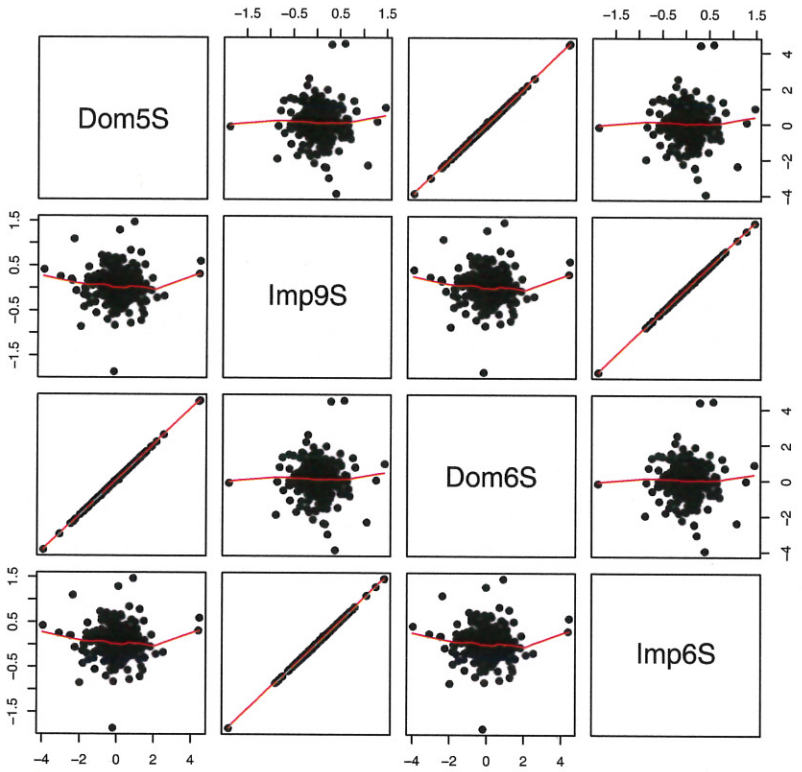


Figure 4.4: Paired scatter plot of estimated dominance and imprinting effects using simulated imprinting data. Dom5S and Dom6S are estimated dominance effects using Models 4.5 and 4.6, respectively; Imp9S and Imp6S are estimated imprinting effects using Models 4.9 and 4.6, respectively.

as expected. As a comparison, the smallest p-value obtained when testing for imprinting using the Dom data was 0.0043, ranked only in the 9^{th} place if the Imp data was used. Considering that there were only 10 loci with a true imprinting effect in the simulation, a locus with p-value ranked in the 9^{th} place would not be detected if the significance threshold was set appropriately. Therefore, it seemed unlikely that a locus would be spuriously claimed as "imprinted" if the true imprinting was absent. Also, the existence of imprinting did not have a large impact on detecting an additive effect, since the detected additive loci using either Dom or Imp data were largely overlapping (Figure 4.6).

Through simulation, we corroborated that, in general, imprinting would not be erroneously claimed if it does not exist and would not be confounded by dominance in a population under

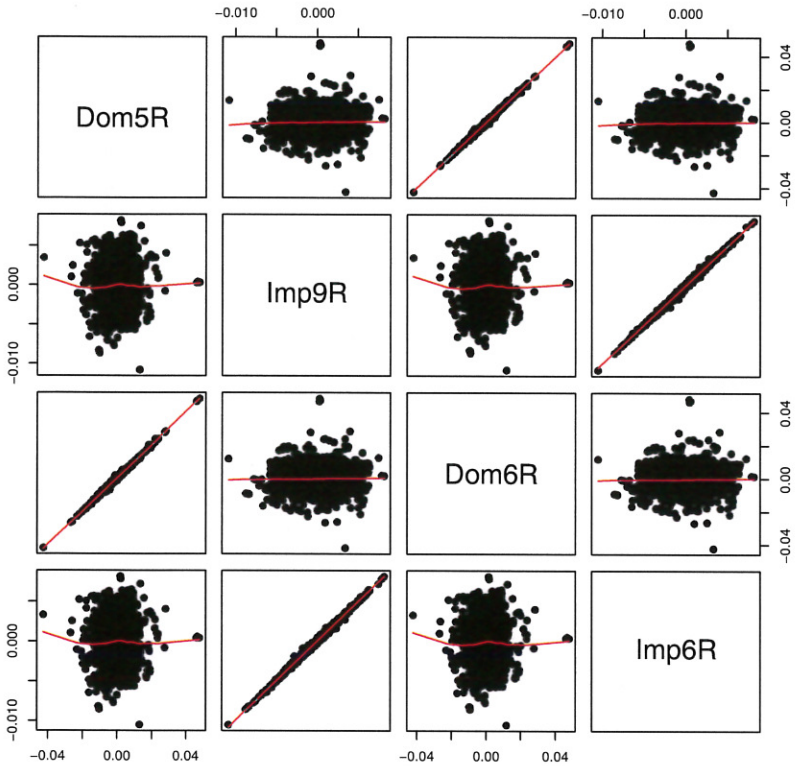


Figure 4.5: Paired scatter plot of estimated dominance and imprinting effects using real mouse data. Dom5R and Dom6R are estimated dominance effects using Models 4.5 and 4.6, respectively; Imp9R and Imp6R are estimated imprinting effects using Models 4.9 and 4.6, respectively.

Hardy-Weinberg equilibrium. Therefore, it is likely that the higher estimate of marker-explained variance in the mouse population was indeed due to imprinting. Hence, the failure of capturing variation attributed to existing imprinting using an additive model may lead to an underestimate of marked variance.

4.4.4 Elevated Marked Variance – Just Because of More Markers?

We found that incorporating genomic imprinting in GWAS produced a larger estimate of phenotypic variance accounted for by significant markers. However, when we estimated the marked variance under imprinting, additively significant markers were also included. Therefore, one may argue that,

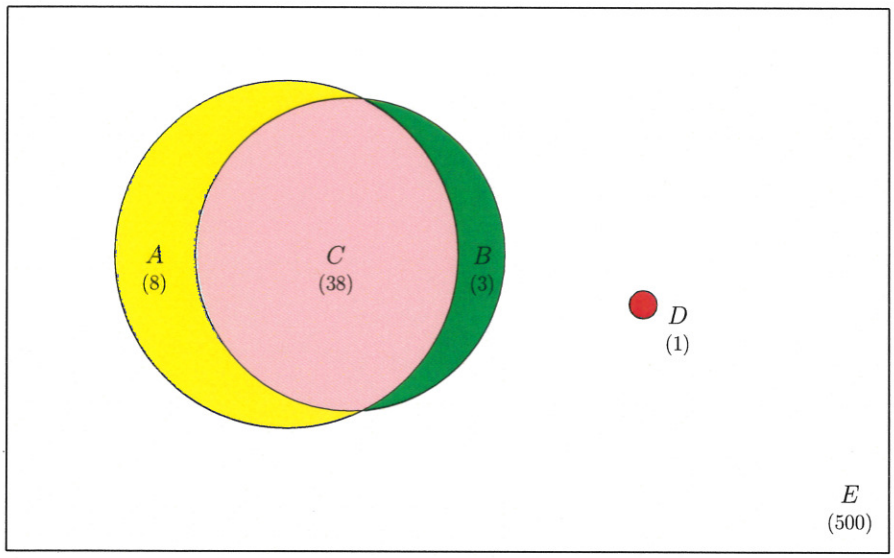


Figure 4.6: Venn's diagram illustrating simulation results. The total number of loci was 500 (E) and 100 had a true effect. 46 loci were detected as additive (A plus C) using the Dom data, and 41 were found (B plus C) using the Imp data. The same model testing procedure found 38 additive loci in common (C) using two data sets, indicating that imprinting does not have a big impact on detecting additive loci. One locus was found to be imprinted (D) using the Imp data and none were detected using the Dom data.

by construction, the marked variance under imprinting would always be larger as long as markers significant on imprinting are detected (Equations 4.7 versus 4.8). Because an objective was to assess how imprinting affects the estimate of marked variance, the following evaluation was considered: we took all markers listed in Table 4.1, but instead of using a "correct" model, we used a "wrong" model to re-estimate the variance accounted for by these markers, i.e., if one marker was detected as imprinted, we now use an additive model to estimate the marked variance, and vice versa for the additively significant markers. The results of this procedure are in Table 4.3.

When markers are imprinted, as indicated in Table 4.1, "erroneously" using an additive model produced a much lower estimate of marked variance (decreased by 85%). This is because the variance of a locus under the imprinting model is $2pq(a+(q-p)d)^2+2pqi^2$ (Equation 4.2), where the term $2pqi^2$ ($\neq 0$ if imprinting exists) is due to imprinting. If an additive model was used, this term would be lost, producing a lower estimate of marked variance. If a marker shows additivity

Table 4.3: Marked variance when marker effects were estimated using the "correct" (boldfaced) and the "wrong" models.

Detected marker (No.)	Model used to estimate marker effects	Marked variance ($\times 10^{-4}$)
Additive (3)	**Additive**	**1.218**
	Imprinting	1.231
Imprinted (6)	Additive	0.091
	Imprinting	**0.624**

but not imprinting, on the other hand, it would be expected that $i = 0$ and hence applying an imprinting model on an additive marker would not increase the variance. However, in practice, it is unlikely that the estimate of i is exactly zero, and hence using a "wrong" model on an additive maker may give a slightly higher variance. However, the difference is negligible (1.218×10^{-4} versus 1.231×10^{-4}, Table 4.3).

4.4.5 Can Imprinting Explain Part of Missing Heritability?

As stated before, heritability is an important parameter in genetic analysis and is usually taken as $2pq\alpha^2/\sigma_P^2$ at a single locus. In GWAS, summation of $2pq\alpha^2$ across all significant markers gives the total marked additive genetic variance under the assumption of linkage equilibrium between markers. The ratio between this total marked variance and phenotypic variance σ_P^2 is usually referred to as the "bottom up" heritability in GWAS (Zaitlen and Kraft, 2012; Zuk *et al.*, 2012). However, marked additive genetic variance differs from additive genetic variance (e.g., Gianola *et al.*, 2009; de los Campos *et al.*, 2015). Therefore, one needs to be cautious when interpreting this "marked" variance, and it is often observed in GWAS that heritability estimated using only statistically significant markers is much lower than pedigree derived heritability, termed as "top-down" estimate in some literature (e.g., Zaitlen and Kraft, 2012). This issue is commonly known as the "missing heritability" problem (Maher, 2008). For example, human height is a trait with estimates of heritability from family studies as high as 0.8 (Silventoinen *et al.*, 2003; Macgregor

et al., 2006), but the variation captured by significant SNP markers from GWAS may take only a proportion of 5∼10% of the total (Weedon *et al.*, 2008; Visscher, 2008; Lango Allen *et al.*, 2010).

Finding sources of missing heritability has been a topic of much interest in genetic and epidemiological studies. The most obvious and likely explanation for this phenomenon is that most traits are polygenic and that markers are in incomplete LD with QTLs as illustrated in Wray *et al.* (2013): h^2_M, the proportion of marker-explained variation, is always smaller than h^2, unless the SNP markers can explain all genetic variation due to perfect LD between markers and causal loci (or in rare cases where some SNP markers are the causal loci themselves), in which case $h^2_M = h^2$. Unfortunately, this situation is unlikely to be encountered in practice. Further, if only genome-wide-significant (GWS) markers are used in genetic analysis, the variation captured by the significant markers (h^2_{GWS}) would be even smaller, resulting in a large amount of missing heritability, measured by $h^2 - h^2_{GWS}$ (Zaitlen and Kraft, 2012; Wray *et al.*, 2013). Therefore, a more appropriate approach could be combining information from both significant markers and pedigree that reflects a "major gene model" situation commonly observed in complex traits analysis, where markers represent the major gene part and pedigree represents the infinitesimal part. Further, although much missing heritability can be recovered by simultaneously including all available dense markers in a statistical model (Yang *et al.*, 2010), the upper bound of this improvement is h^2_M, indicating that the variation hidden by incomplete LD relationships between markers and QTLs is "still missing" and difficult to be restored (Wray *et al.*, 2013). The covariance between alleles stemming from LD complicates the variance assessment, and epistatic effects, i.e., interactions between causal loci are often ignored. These two issues can also lead to dubious attributions of genetic variation (Gibson, 2010; Gianola *et al.*, 2013).

Epigenetic variation has been suggested as another potential source of missing heritability (e.g., Manolio *et al.*, 2009; Eichler *et al.*, 2010). From the imprinting model introduced above, it is expected that imprinting may have an impact on additive genetic variance of a single locus, and hence affect the bottom up estimate of heritability, as evidenced in our analysis. Combining the

bottom-up and top-down genetic variation may lead to a less clear result since infinitesimal effects contributed more to the additive genetic variance than markers, and the estimates of this component were close to each other when using two approaches (Table 4.2). Nevertheless, incorporating imprinting still resulted in a 10% increase on heritability, as the estimates with and without consideration of imprinting using variance components in Table 4.2 are

$$\hat{h}^2_{\text{Imp}} = \frac{1.842 + 3.816}{1.842 + 3.816 + 4.742 + 18.687} = 0.195,$$

and

$$\hat{h}^2_{\text{Men}} = \frac{1.218 + 3.805}{1.218 + 3.805 + 4.747 + 18.694} = 0.176,$$

respectively. Thus, a higher estimate of additive variability was found whether the pedigree information was included or not. This result indicated that the underestimate of additive variation by erroneously using an additive model in the case of imprinting could be a potential source of missing heritability in GWAS, as discussed in Manolio *et al.* (2009) and Eichler *et al.* (2010).

4.4.6 Imprinting Effect and Parent-of-Origin Effect

Our results indicated that existing imprinting effects should not be ignored in genetic analysis. Meanwhile, it is also important to make a distinction between the terms "imprinting effect" and "parent-of-origin effect". These terms are often used exchangeably in much of the epigenetic literature. However, a parent-of-origin effect referring to different genetic contribution of different parents to offspring is a broader concept than an imprinting effect. Genomic imprinting is the most important source of parent-of-origin effects, but not the only one. For example, maternal effects observed in swine production is a well known form of parent-of-origin effect that is not known to involve any epigenetic mechanisms; reciprocal effects observed in poultry breeding is another type of parent-of-origin effect. In the imprinting model that was adopted in our analysis, all inferences

were performed at the DNA level using SNP markers. Hence, not all "putatively" imprinted loci were necessarily caused by imprinting at the epigenetic level. Moreover, other factors can lead to the detection of spurious imprinting effect that is actually caused by other types of parent-of-origin effect (e.g., Hager *et al.*, 2008; Tuiskula-Haavisto *et al.*, 2004; Tuiskula-Haavisto and Vilkki, 2007) or even by linkage disequilibrium between markers (Sandor and Georges, 2008). Therefore, results from this study should be viewed as parent-of-origin effects instead of imprinting. On the other hand, if the objective of a certain study is to determine or verify imprinting status, we recommend that examination of variation must be taken place at the epigenetic level using, for instance, differential methylation analysis. However, this does not contradict the statement that an underestimate of additive variability would occur if existing imprinting was ignored.

4.5 Conclusion

We were inspired by studies that proposed equivalent one-locus imprinting models for quantitative genetic analysis. These studies defined paternal and maternal gene substitution effects. As such, imprinting does contribute to additive variance and a partition of additive variance into unimprinted and imprinted components is available. This variance decomposition hints that heritable genetic variation induced by epigenetic mechanisms, especially genomic imprinting, may have a considerable impact on the underlying genetic architecture of some complex traits, but it is largely neglected in many studies. Specifically, narrow sense heritability, especially marked "bottom up" heritability in GWAS, may be underestimated if one ignores imprinting when it is present. We tested this using a genome-wide association study performed on mouse BMI data. Results indicated that the portion of phenotypic variation explained by significant SNP markers increased drastically when imprinting effects were considered.

Moreover, the imprinting regression model used here detects differences between paternally and maternally inherited alleles regardless of whether the biological mechanism is imprinting or not.

Hence, this model might be capturing other (either genetic or epigenetic) mechanisms that produce nonequivalent contributions of paternal and maternal genomes as well. Therefore, it would be more appropriate to refer to this model as a model incorporating parent-of-origin effects, and such that, it can be applied to a number of situations. For example, in the human genome, more than 50% of the genes have shown preferential expression of the paternal or maternal allele due to various mechanisms Lo *et al.* (2003), indicating that our approach may apply to a wide range of complex traits, whenever reciprocal heterozygotes generate different genotypic values. Since imprinting is only one of such mechanisms, it is possible that more (epigenetic) sources of phenotypic variation and of missing heritability may be uncovered in the future. Nevertheless, imprinting is widely considered as the most important source of parent-of-origin effects, so in order to avoid a possibly wrong inference on genetic architecture of a complex trait of interest, imprinting should not be neglected if indication of the presence of imprinting exists.

References

Bates, D. and A. I. Vázquez, 2013. *pedigreemm: Pedigree-based mixed-effects models.* R package version 0.2-4

Browning, B. L., 2011. *BEAGLE 3.3.2 User's Manual*

Browning, B. L. and S. R. Browning, 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet., 84(2): 210–223

Cheverud, J. M., R. Hager, C. Roseman, *et al.*, 2008. Genomic imprinting effects on adult body composition in mice. Proc. Natl. Acad. Sci. U.S.A., 105(11): 4253–4258

Coster, A., O. Madsen, H. C. Heuven, *et al.*, 2012. The imprinted gene *DIO3* is a candidate gene for litter size in pigs. PLoS ONE, 7(2): e31825

de los Campos, G., D. Sorensen, and D. Gianola, 2015. Genomic Heritability: What Is It? PLoS Genet., 11(5): e1005048

Dong, C., W. D. Li, F. Geller, *et al.*, 2005. Possible genomic imprinting of three human obesity-related genetic loci. Am. J. Hum. Genet., 76(3): 427–437

Eichler, E. E., J. Flint, G. Gibson, *et al.*, 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet., 11(6): 446–450

Falconer, D. S. and T. F. C. Mackay, 1996. *Introduction to Quantitative Genetics.* Prentice Hall, 4th edition

Fisher, R. A., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb., 52: 399–433

Georges, M., C. Charlier, and N. Cockett, 2003. The callipyge locus: evidence for the *trans* interaction of reciprocally imprinted genes. Trends Genet., 19(5): 248–252

Gianola, D., G. de los Campos, W. G. Hill, *et al.*, 2009. Additive genetic variability and the Bayesian alphabet. Genetics, 183(1): 347–363

Gianola, D., F. Hospital, and E. Verrier, 2013. Contribution of an additive locus to genetic variance when inheritance is multi-factorial with implications on interpretation of GWAS. Theor. Appl. Genet., 126(6): 1457–1472

Gibson, G., 2010. Hints of hidden heritability in GWAS. Nat. Genet., 42(7): 558–560

Gorlova, O. Y., C. I. Amos, N. W. Wang, *et al.*, 2003. Genetic linkage and imprinting effects on body mass index in children and young adults. Eur. J. Hum. Genet., 11(6): 425–432

Hager, R., J. M. Cheverud, and J. B. Wolf, 2008. Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. Genetics, 178(3): 1755–1762

Henderson, C. R., 1984. *Applications of Linear Models in Animal Breeding.* University of Guelph, Ontario, Canada

Jerez-Timaure, N. C., E. J. Eisen, and D. Pomp, 2005. Fine mapping of a QTL region with large effects on growth and fatness on mouse chromosome 2. Physiol. Genomics, 21(3): 411–422

Kempthorne, O., 1954. The correlation between relatives in a random mating population. Proc. R. Soc. Lond., B, Biol. Sci., 143(910): 103–113

Kilpinen, H. and E. T. Dermitzakis, 2012. Genetic and epigenetic contribution to complex traits. Hum. Mol. Genet., 21(R1): R24–28

Kobayashi, M., F. Io, T. Kawai, *et al.*, 2006. Major quantitative trait locus on chromosome 2 for glucose tolerance in diabetic SMXA-5 mouse established from non-diabetic SM/J and A/J strains. Diabetologia, 49(3): 486–495

Lango Allen, H., K. Estrada, G. Lettre, *et al.*, 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature, 467(7317): 832–838

Lawson, H. A., J. M. Cheverud, and J. B. Wolf, 2013. Genomic imprinting and parent-of-origin effects on complex traits. Nat. Rev. Genet., 14(9): 609–617

Leamy, L. J., D. Pomp, E. J. Eisen, *et al.*, 2002. Pleiotropy of quantitative trait loci for organ weights and limb bone lengths in mice. Physiol. Genomics, 10(1): 21–29

Lo, H. S., Z. Wang, Y. Hu, *et al.*, 2003. Allelic variation in gene expression is common in the human

genome. Genome Res., 13(8): 1855–1862

Luedi, P. P., A. J. Hartemink, and R. L. Jirtle, 2005. Genome-wide prediction of imprinted murine genes. Genome Res., 15(6): 875–884

Lynch, M. and B. Walsh, 1998. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates

Macgregor, S., B. K. Cornes, N. G. Martin, *et al.*, 2006. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. Hum. Genet., 120(4): 571–580

Maher, B., 2008. The case of the missing heritability. Nature, 456(7218): 18–21

Manolio, T. A., F. S. Collins, N. J. Cox, *et al.*, 2009. Finding the missing heritability of complex diseases. Nature, 461(7265): 747–753

Meijers-Heijboer, E. J., L. A. Sandkuijl, H. G. Brunner, *et al.*, 1992. Linkage analysis with chromosome 15q11-13 markers shows genomic imprinting in familial Angelman syndrome. J. Med. Genet., 29(12): 853–857

Moskvina, V. and K. M. Schmidt, 2008. On multiple-testing correction in genome-wide association studies. Genet. Epidemiol., 32(6): 567–573

Neugebauer, N., I. Räder, H. J. Schild, *et al.*, 2010. Evidence for parent-of-origin effects on genetic variability of beef traits. J. Anim. Sci., 88(2): 523–532

Nicholls, R. D., S. Saitoh, and B. Horsthemke, 1998. Imprinting in Prader-Willi and Angelman syndromes. Trends Genet., 14(5): 194–200

Rance, K. A., J. M. Fustin, G. Dalgleish, *et al.*, 2005. A paternally imprinted QTL for mature body mass on mouse chromosome 8. Mamm. Genome, 16(8): 567–577

Reik, W. and J. Walter, 2001. Genomic imprinting: parental influence on the genome. Nat. Rev. Genet., 2(1): 21–32

Rocha, J. L., E. J. Eisen, L. D. Van Vleck, *et al.*, 2004. A large-sample QTL study in mice: I. Growth. Mamm. Genome, 15(2): 83–99

Sandell, L. L., X. J. Guan, R. Ingram, *et al.*, 2003. Gatm, a creatine synthesis enzyme, is imprinted in mouse placenta. Proc. Natl. Acad. Sci. U.S.A., 100(8): 4622–4627

Sandor, C. and M. Georges, 2008. On the detection of imprinted quantitative trait loci in line crosses: effect of linkage disequilibrium. Genetics, 180(2): 1167–1175

Searle, S., 1971. *Linear Models.* John Wiley & Sons, New Jersey

Searle, S., G. Casella, and C. McCulloch, 2006. *Variance Components.* John Wiley & Sons, New Jersey

Silventoinen, K., S. Sammalisto, M. Perola, *et al.*, 2003. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res., 6(5): 399–408

Sorensen, D. and D. Gianola, 2002. *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. Springer, New York

Spencer, H. G., 2002. The correlation between relatives on the supposition of genomic imprinting. Genetics, 161(1): 411–417

Spencer, H. G., 2009. Effects of genomic imprinting on quantitative traits. Genetica, 136(2): 285–293

Stoehr, J. P., S. T. Nadler, K. L. Schueler, *et al.*, 2000. Genetic obesity unmasks nonlinear interactions between murine type 2 diabetes susceptibility loci. Diabetes, 49(11): 1946–1954

Tuiskula-Haavisto, M., D. J. de Koning, M. Honkatukia, *et al.*, 2004. Quantitative trait loci with parent-of-origin effects in chicken. Genet. Res., 84(1): 57–66

Tuiskula-Haavisto, M. and J. Vilkki, 2007. Parent-of-origin specific QTL – a possibility towards understanding reciprocal effects in chicken and the origin of imprinting. Cytogenet. Genome Res., 117(1-4): 305–312

Valdar, W., L. C. Solberg, D. Gauguier, *et al.*, 2006. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat. Genet., 38(8): 879–887

Vázquez, A. I., D. Bates, G. J. Rosa, *et al.*, 2010. Technical note: an R package for fitting generalized linear mixed models in animal breeding. J. Anim. Sci., 88(2): 497–504

Visscher, P. M., 2008. Sizing up human height variation. Nat. Genet., 40(5): 489–490

Wang, Z., Z. Wang, J. Wang, *et al.*, 2012. A quantitative genetic and epigenetic model of complex traits. BMC Bioinformatics, 13: 274

Warnes, G., with contributions from Gregor Gorjanc, F. Leisch, *et al.*, 2012. *genetics: Population Genetics*. R package version 1.3.8

Weedon, M. N., H. Lango, C. M. Lindgren, *et al.*, 2008. Genome-wide association analysis identifies 20 loci that influence adult height. Nat. Genet., 40(5): 575–583

Williamson, C. M., A. Blake, S. Thomas, *et al.*, 2013. Mouse Imprinting Data and References. MRC Harwell, Oxfordshire. World Wide Web Site `http://www.har.mrc.ac.uk/research/genomic_imprinting/`

Wolf, J. B., J. M. Cheverud, C. Roseman, *et al.*, 2008a. Genome-wide analysis reveals a complex pattern of genomic imprinting in mice. PLoS Genet., 4(6): e1000091

Wolf, J. B., R. Hager, and J. M. Cheverud, 2008b. Genomic imprinting effects on complex traits: a phenotype-based perspective. Epigenetics, 3(6): 295–299

Wood, A. J., R. Schulz, K. Woodfine, *et al.*, 2008. Regulation of alternative polyadenylation by genomic imprinting. Genes Dev., 22(9): 1141–1146

Wray, N. R., J. Yang, B. J. Hayes, *et al.*, 2013. Pitfalls of predicting complex traits from SNPs.

Nat. Rev. Genet., 14(7): 507–515

Yang, J., B. Benyamin, B. P. McEvoy, *et al.*, 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet., 42(7): 565–569

Zaitlen, N. and P. Kraft, 2012. Heritability in the genome-wide association era. Hum. Genet., 131(10): 1655–1664

Zuk, O., E. Hechter, S. R. Sunyaev, *et al.*, 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. U.S.A., 109(4): 1193–1198

Chapter 5

Incorporating Parent-of-Origin Effects in Whole-Genome Prediction of Complex Trait

Parent-of-origin effects are due to differential contributions of paternal and maternal lineages to offspring phenotypes. Epigenetically induced parent-of-origin effects have attracted attention recently due to a potential impact on variation of complex traits. Given that prediction of genetic merit or phenotypic performance is of interest in complex traits studies, it is relevant to consider parent-of-origin effects in such predictions. In this chapter, a whole-genome prediction model incorporating parent-of-origin effects was built by considering parental allelic substitution effects of SNP markers and gametic relationships derived from a pedigree. Predictive performance of this model was compared with that of an additive model without parent-of-origin effects using real and simulated data.

5.1 Introduction

Parent-of-origin effects are asymmetric influences exerted on offspring, based on the sex of the parent. Genomic imprinting, differential and/or preferential gene expression usually caused by differential DNA methylation (Li *et al.*, 1993; Delaval and Feil, 2004) or histone modification (McEwen and Ferguson-Smith, 2009) on different parental alleles, is one of the most studied epigenetic mechanisms and an important cause of parent-of-origin effects. Imprinting has impact on several human diseases (Hall, 1990; Solter, 1992; Falls *et al.*, 1999; Clayton-Smith, 2003; Úbeda and Wilkins, 2008) such as the Prader-Willi (PWS) and Angelman (AS) syndromes (Meijers-Heijboer *et al.*, 1992; Nicholls *et al.*, 1998), as well as complex traits in livestock (Wolf *et al.*, 2008b; Kilpinen and Dermitzakis, 2012; Lawson *et al.*, 2013). For example, *i*QTL mapping studies have detected imprinted QTLs affecting economically important traits in swine (Knott *et al.*, 1998; Jeon *et al.*, 1999; Nezer *et al.*, 1999; de Koning *et al.*, 2002; Stella *et al.*, 2003; Lee *et al.*, 2003; Thomsen *et al.*, 2004; Kim *et al.*, 2007), beef cattle (Engellandt and Tier, 2002; Essl and Voith, 2002; Meyer and Tier, 2012), sheep (Lewis and Redrup, 2005), mice (Cui *et al.*, 2006, 2007), and dogs (Liu *et al.*, 2007). In addition, genome wide scan studies with dense marker chips have also suggested imprinted loci associated with complex traits in various mammalian species (e.g., Holl *et al.*, 2004; Wolf *et al.*, 2008a; Cheverud *et al.*, 2008; Imumorin *et al.*, 2011; Kärst *et al.*, 2012; Coster *et al.*, 2012).

QTL mapping studies can flag genomic regions that contribute to traits of interest and to marker assisted selection (MAS, Ribaut and Hoisington, 1998; Guimarães *et al.*, 2007). However, use of QTL mapping for breeding purposes has failed to yield clear dividends (e.g., Young, 1999; Dekkers, 2004). A possible explanation is that QTL mapping studies require, e.g., using carefully designed cross breeding experiments and this material is seldom available in livestock. Thus, artificial selection using predicted genetic merit of candidate individuals still dominates animal improvement programs. Breeding values have been predicted from resemblance between relatives using pedigree information (e.g., Henderson, 1984; Mrode, 2014). In the genomics era, however, the availability

of high throughput genotyping techniques makes it possible to investigate genotypes of hundreds of thousands or even millions of SNP (single nucleotide polymorphism) markers simultaneously, resulting in what is known as "genomic selection" or "whole-genome prediction" (Meuwissen *et al.*, 2001; Goddard, 2009; de los Campos *et al.*, 2013). With continuously decreasing genotyping costs, genomic selection has become affordable for commercial settings in some species (Schaeffer, 2006), and QTL mapping is not used much in animal breeding unless the objective is to find a major gene. Even in crops, genomic selection is replacing QTL-MAS gradually. Although some debate persists (Jonas and de Koning, 2013), genomic selection will probably dominate in the foreseeable future (Nakaya and Isobe, 2012).

Genomic selection (GS) and whole-genome prediction (WGP) are performed by exploiting associations between phenotypes and an enormous number of SNP markers under certain assumptions on the underlying genetic architecture. Often, the association between phenotype and SNPs is explored using markers as covariates in a linear regression model. Since the number of covariates (p) is usually much larger than the number of observations (n), different techniques have been employed to circumvent the "curse of dimensionality" in GS/WGP studies. Commonly used methods include Bayesian regression (e.g., Meuwissen *et al.*, 2001; Gianola *et al.*, 2009; de los Campos *et al.*, 2013; Gianola, 2013), G-BLUP (e.g., VanRaden, 2008; Legarra *et al.*, 2009), semi-parametric methods (e.g., Gianola *et al.*, 2006; Gianola and de los Campos, 2008; de los Campos *et al.*, 2009a; Morota and Gianola, 2014) and neural networks (e.g., Gianola *et al.*, 2011; González-Camacho *et al.*, 2012; Pérez-Rodríguez *et al.*, 2012), among others. All these models have assumed that the inheritance of the complex trait is Mendelian, i.e., paternally and maternally inherited allele are functionally equivalent. Under this assumption, no phenotypic difference between genotypes A_1A_2 and A_2A_1 is expected. Markers are assigned codes such as 0, 1 or 2 according to genotype at the locus, and the average substitution effects of all markers in the model are estimated simultaneously. Prediction is then performed by combining these marker effect estimates with a genotype matrix in an independent set of individuals. However, recent studies suggest that some traits are not strictly

Mendelian. For example, Mott *et al.* (2014) found that 91 out of 97 murine traits were subject to parent-of-origin effects. In a review, Lawson *et al.* (2013) also suggested that parent-of-origin effects may be more prevalent than previously thought. Perhaps parent-of-origin effects may enhance WGP models, if considered appropriately.

Currently used GS models may not be suitable for parent-of-origin-effects-affected traits, where inheriting one allele from the father may have a different effect on phenotype than when the same allele is from the mother. This suggests that two distinct substitution effects associated with the two parental origins of an allele are needed. A one-locus quantitative genetic model with consideration of imprinting has been proposed (Spencer, 2002; Shete and Amos, 2002; de Koning *et al.*, 2002), where genotypes A_2A_2, A_1A_2, A_2A_1 and A_1A_1 are assumed to have genotypic values $-a$, $d-i$, $d+i$ and a, respectively, and paternal and maternal allelic substitution effects are defined as $\alpha_{♂} = a + d(q-p) + i$ and $\alpha_{♀} = a + d(q-p) - i$. In Chapter 4, a GWAS-like scan conducted with this model indicated that ignoring existing imprinting may underestimate additive genetic variation. Here, we extend this model by including all available markers simultaneously into a prediction model and assess whether or not this model improves prediction of phenotypes over the additive model currently employed in WGP.

Before proceeding, some clarification is necessary. In much of the epigenetic literature, the terms "imprinting effects" and "parent-of-origin effects" have been used interchangeably. For example, in "*i*QTL mapping" the detected QTLs are putatively *i*mprinted. However, the statistical model used in *i*QTL mapping does not guarantee that the detected parent-of-origin effects are necessarily due to imprinting. A counter-example was presented by Hager *et al.* (2008) where maternal effects can mimic imprinting effects in statistical analysis. Furthermore, parent-of-origin effects were detected in birds (Tuiskula-Haavisto *et al.*, 2004; Tuiskula-Haavisto and Vilkki, 2007), although no strong evidence of imprinting in birds is available (O'Neill *et al.*, 2000; Nolan *et al.*, 2001; Frésard *et al.*, 2013). Thus, results obtained from the model described above and its variants should be interpreted as parent-of-origin effects but not beyond (Chapter 4). In this study, we

build WGP models to incorporate parent-of-origin effects, aiming at obtaining a higher predictive accuracy when a complex trait is subject to parent-of-origin effects. We use the term "parent-of-origin effects" throughout, but in simulations we mimicked imprinting mechanisms to simplify the source of parent-of-origin effects. The simulated data was used for model evaluation under various conditions.

This chapter is organized as follows. First, a previously proposed mixed effect model incorporating parent-of-origin effects at the lineage level is introduced. Then, the one-locus imprinting model discussed in Chapters 3 and 4 is extended to incorporate all available markers simultaneously to include parent-of-origin effects at the DNA (SNP markers) level, and our prediction model is constructed using both pedigree and DNA information. This model is applied to real (mouse) and simulated data, and the predictive performance is compared to that from an additive model. Following a section arguing advantages and drawbacks of this model, a discussion on possibilities and challenges of conducting whole genome prediction using epigenetic information to incorporate parent-of-origin effects is provided.

5.2 Prediction Model Incorporating Parent-of-Origin Effects

Consider the pedigree-based additive effects model (e.g., Henderson, 1984; Mrode, 2014):

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{5.1}$$

where the $n \times 1$ vector $\boldsymbol{y}$ contains phenotypic records; μ is an effect common to all individuals; $\boldsymbol{b}$ is a vector of fixed effects with associated incidence matrix $\boldsymbol{X}$; $\boldsymbol{u}$ is the $n \times 1$ vector of normally distributed infinitesimal additive effects with zero mean vector and variance-covariance matrix $\boldsymbol{A}\sigma_A^2$, where $\boldsymbol{A}$ is the $n \times n$ pedigree-based numerator relationship matrix and σ_A^2 is the additive genetic variance; and $\boldsymbol{e}$ is the residual vector whose elements are assumed to be independent and identically distributed as normal with zero mean and variance σ_e^2. A commonly used technique for making

predictions of yet-to-be-observed data is best linear unbiased prediction (BLUP, Henderson, 1984), where estimation of $\boldsymbol{b}$ and prediction of $\boldsymbol{u}$ are performed simultaneously. Variance components can be estimated, for example, by restricted maximum likelihood (REML).

If dense markers (e.g., SNPs) are available, the following model can be used for genome-enabled prediction (e.g., Meuwissen *et al.*, 2001):

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \sum_{j=1}^{p} \boldsymbol{w}_j \alpha_j + \boldsymbol{e}. \tag{5.2}$$

Here, p is the (possibly large) number of markers and the assumption is that QTL contributing to the phenotype $\boldsymbol{y}$ are in linkage disequlibrium (LD) with at least one SNP marker. In this model, α_j is the substitution effect of the j^{th} marker with genotype code W_{ij} ($W_{ij} = 0$, 1 or 2 for genotypes A_2A_2, A_1A_2/A_2A_1 or A_1A_1) for the i^{th} individual, and $\{\sum_{j=1}^{p} \boldsymbol{w}_j \alpha_j\} = \boldsymbol{W}\boldsymbol{\alpha}$, where $\boldsymbol{W}$ is $n \times p$, $\boldsymbol{\alpha}$ is $p \times 1$, and $\boldsymbol{w}_j$ is the j^{th} column of $\boldsymbol{W}$. Marker effects can be learned Bayesianly (e.g., Meuwissen *et al.*, 2001; Gianola *et al.*, 2009) by drawing samples from posterior distributions using Markov Chain Monte Carlo (MCMC) techniques. Predictive performance using Model 5.2 is often better than Model 5.1, and several studies suggested that including both pedigree and marker information can achieve an even higher prediction accuracy (de los Campos *et al.*, 2009b; Erbe *et al.*, 2010).

Models described above assume that QTL and markers are inherited in a Mendelian manner. However, in the presence of imprinting, or more generally, parent-of-origin effects, receiving one allele from the mother might have a different effect on $\boldsymbol{y}$ than receiving the same allele from the father (Reik and Walter, 2001; Spencer, 2002; Shete and Amos, 2002). Before the genomic era, the following mixed model using pedigree information was proposed to account for parent-of-origin effects (Gibson *et al.*, 1988; Schaeffer *et al.*, 1989):

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{M}\boldsymbol{g} + \boldsymbol{e}, \tag{5.3}$$

where $\boldsymbol{y}$, μ, $\boldsymbol{b}$, $\boldsymbol{u}$ and $\boldsymbol{e}$ are as in Model 5.1; $\boldsymbol{g}$ is a $2n \times 1$ vector of additional genetic effects expressed

only when inherited from a maternal or paternal gamete assumed that $\boldsymbol{g} \sim N(\boldsymbol{0}, \boldsymbol{L}\sigma_g^2)$ with $\boldsymbol{L}$ being a $2n \times 2n$ gametic relationship matrix calculated from a known pedigree. When considering SNP markers, the one-locus regression model proposed by Shete and Amos (2002) (Model 3.1, Chapter 3) can be extended to include all available markers simultaneously as in WGP studies. Thus, we combined Models 5.3 and 3.1, which contained both pedigree and marker information, into a WGP model (referred to as POE model hereafter) suitable for traits affected by parent-of-origin effects:

$$\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{M}\boldsymbol{g} + \sum_{j=1}^{p} \boldsymbol{I}_{j♀}\alpha_{j♀} + \sum_{j=1}^{p} \boldsymbol{I}_{j♂}\alpha_{j♂} + \boldsymbol{e}, \tag{5.4}$$

where $\alpha_♀$ and $\alpha_♂$ are the effects of receiving an A_1 allele from the female and male parents (maternal and paternal allelic substitution effects), respectively, and $\boldsymbol{I}_♀$ and $\boldsymbol{I}_♂$ are vectors of associated indicator variables that can take values 0 or 1. To evaluate the performance of the POE model (5.4), it was compared with an additive model (referred to as ADD model hereafter) without parent-of-origin effects at either the pedigree or marker levels. Model ADD is then:

$$\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \sum_{j=1}^{p} \boldsymbol{W}_j\alpha_j + \boldsymbol{e}. \tag{5.5}$$

5.3 Materials and Model Evaluation Scenarios

5.3.1 Data

The mouse data used in Chapter 4 was also used here. All fixed and random effects were the same as in Chapter 4 and models POE and ADD (as in Equations 5.4 and 5.5, respectively) were fitted to perform whole genome predictions on body mass index (BMI). The additive relationship matrix $\boldsymbol{A}$ and the gametic relationship matrix $\boldsymbol{L}$ were calculated from pedigree using the R package synbreed (Wimmer *et al.*, 2012). We also used simulated data to evaluate the performance of the POE and of the ADD models under different situations. Parent-of-origin effects was simulated in the two-step

procedure described below.

First, we used QMSim (Sargolzaei and Schenkel, 2009) to simulate a genome of 10 pairs of chromosomes each of length 1 Morgan. Each chromosome had 1,000 randomly located bi-allelic SNP markers, so there were ten thousands markers in total, as in the mouse data. Approximately 150 simulated QTLs were randomly located in the genome and these were not chosen from the simulated SNP markers. QTL effects were randomly drawn from a normal distribution with zero mean and variance set to the software default value. The population started from 100 males and 100 females with 1,000 generations of random mating to create LD between QTL and between marker loci with QTL; mutation rates were $u_{\text{QTL}} = 10^{-4}$ per QTL locus and $u_{\text{Mrk}} = 10^{-2}$ per marker locus, respectively. All QTLs and markers were fixed in generation 1. In the three most recent generations, without mutation, the population was expanded into 2,000 individuals per generation with a 1:1 sex ratio.

In step 2, we simulated parent-of-origin effects that mimic imprinting from the raw output of the previous step. For a long time, imprinting has been viewed as a "full-null" phenomenon where the silencing of the imprinted allele is complete whereas the expression of the allele inherited from the other parent is intact (Reik and Walter, 2001). However, genomic imprinting can potentially operate at any level of gene regulation (e.g., at promoters, enhancers, splicing junctions, or polyadenylation sites, etc.) to present a more complex pattern of parent-specific differential expression (Barlow and Bartolomei, 2014). For example, recent studies have provided evidence that, for some imprinted loci, both alleles are differentially expressed in a parent-of-origin-preferential or parent-of-origin-dependent manner (Khatib, 2007), indicating that the silencing is incomplete (Abramowitz and Bartolomei, 2009; Barlow, 2011). Such deviation from canonical imprinting was defined as partial imprinting (Wolf *et al.*, 2008a; Morcos *et al.*, 2011), incorporated in a one-locus imprinting model (Spencer, 2002; Shete and Amos, 2002; de Koning *et al.*, 2002), and also considered in our simulation. Let θ_{ij1} and θ_{ij2} (given by QMSim output) be the two allelic effects of QTL j in individual i obtained from a certain QMSim run. Because QMSim records the parental

origin of these two alleles, θ_{ij1} and θ_{ij2} can be represented by, say, $\theta_{ij♀}$ and $\theta_{ij♂}$, respectively. If this QTL is maternally imprinted, the genotypic value at this QTL for individual i can be written as

$$\rho \cdot \theta_{ij♀} + \theta_{ij♂}, \tag{5.6}$$

where ρ is a parameter that controls the level of imprinting. Five different values were assigned to ρ: 0, 0.25, 0.5, 0.75, and 1, where $\rho = 1$ corresponded to no imprinting, $\rho = 0$ corresponded to complete imprinting, and $\rho = 0.25, 0.5, 0.75$ defined different levels of partial imprinting. We further assumed that a proportion $s = \{0.15, 0.3, 0.45, 0.6\}$ of n_{QTL} QTLs were either paternally or maternally imprinted with equal frequency (a validation on the choice of these values is given in Section 5.5). Hence, the phenotypic value of individual i is

$$y_i = \sum_{j \in \text{NI}} (\theta_{ij♀} + \theta_{ij♂}) + \sum_{j \in \text{MI}} (\rho \cdot \theta_{ij♀} + \theta_{ij♂}) + \sum_{j \in \text{PI}} (\theta_{ij♀} + \rho \cdot \theta_{ij♂}) + \varepsilon_i, \tag{5.7}$$

where NI, MI and PI are sets of $(1-s) \cdot n_{\text{QTL}}$ non-imprinted, randomly selected $\frac{1}{2} s \cdot n_{\text{QTL}}$ maternally imprinted and $\frac{1}{2} s \cdot n_{\text{QTL}}$ paternally imprinted QTLs, respectively, and ε_i is the environmental effect on individual i given by QMSim. Note that the environmental effect ε_i was not changed and that a common ρ was shared by all imprinted QTLs in a particular scenario for simplification.

Equation 5.7 was applied to all three recent generations in all 20 combinations of ρ and s. In subsequent analyses, generation 1,002 was the training set and generation 1,003 was the testing set. This whole procedure was replicated 5 times and the the average performance of all replicates was used for model evaluation.

5.3.2 Model Training and Phenotype Prediction

Both the ADD and the POE models were trained Bayesianly with an implementation of MCMC using the R package BGLR (Pérez and de los Campos, 2014; de los Campos and Pérez Rodríguez, 2014). Each chain was run for 60,000 iterations, with the first 10,000 iterations discarded as burn-in

and the rest were thinned by a factor of 10 to use samples that were mildly correlated.

For the ADD model, the conditional prior distribution of the substitution effect of marker j was a normal distribution with zero mean and variance $\tau_j^2\sigma_e^2$, where σ_e^2 came from a scaled inverted χ^2 distribution with scale S_e and degrees of freedom df_e set to default values in package BGLR (de los Campos and Pérez Rodríguez, 2014); τ_j^2 was drawn from an exponential distribution with parameter $\lambda^2/2$. Hyperparameter λ^2 was drawn from a Gamma distribution with shape s and rate r set to default values. This prior creates a double-exponential posterior density for marker effects, given λ, and is referred to as Bayesian Lasso (Park and Casella, 2008; de los Campos *et al.*, 2009b). The infinitesimal additive effects $\boldsymbol{u}$ had a conditional multivariate normal prior $N(\mathbf{0}, \boldsymbol{A}\sigma_u^2)$, where σ_u^2 was drawn from a scaled inverted χ^2 distribution with scale S_u and degrees of freedom df_u set to default values. Similarly, for cage effects, $\boldsymbol{c}|\sigma_c^2 \sim N(\mathbf{0}, \boldsymbol{I}\sigma_c^2)$ and again, the scale S_c and degrees of freedom df_c for the prior of σ_c^2 were set to default values.

For the POE model, prior distributions were similar to those described above, except that two marker effects, the paternal and maternal allelic substitution effects were included for each marker. The extra vector of gametic effects was assumed to have the distribution $\boldsymbol{g}|\sigma_g^2, \boldsymbol{L}, S_g, df_g \propto N(\boldsymbol{g}|\sigma_g^2, \boldsymbol{L}) \cdot \chi^{-2}(\sigma_g^2|S_g, df_g)$. Again, all hyperparameters for the scaled inverted χ^2 distributions were set to package default values.

5.4 Results

5.4.1 Mouse Data Analysis

After data cleaning, 1,869 individuals were randomly partitioned into three disjoint sets and a 3-fold cross validation (CV) was performed. This 3-fold CV was repeated five times for stability assessment. Table 5.1 gives average results over the 5 replications. The ADD model performed slightly better than the POE model when evaluated by different metrics, but the difference was

minimal. We also used all individuals to estimate the variance components associated with each random effect in the two models (Table 5.2). Note that the POE model had the extra gametic effect $\boldsymbol{g}$ which had two variance components associated with pedigree: the infinitesimal additive variance (σ_u^2) and the gametic variance (σ_g^2), contrary to a single component in the ADD model. The estimates $\hat{\sigma}_u^2$, $\hat{\sigma}_c^2$, and $\hat{\sigma}_e^2$ in the POE model were all smaller than these in the ADD model, because of the impact of the extra component in the POE model (i.e., $\hat{\sigma}_g^2$). Pedigree-based heritabilities of BMI estimated from these two models were

$$\hat{h}^2_{\text{ADD}} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2} = 0.104,$$

and

$$\hat{h}^2_{\text{POE}} = \frac{\hat{\sigma}_u^2 + \hat{\sigma}_g^2}{\hat{\sigma}_u^2 + \hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2} = 0.132.$$

This result is in agreement with a recent study showing that considering parent-of-origin effects increases heritability estimates in several mouse traits (Mott *et al.*, 2014), and is also consistent with the previous study where ignoring parent-or-origin effects produced a smaller heritability estimate of BMI (Chapter 4). However, $\hat{h}^2_{\text{ADD}}$ given above was lower than the estimate of Valdar *et al.* (2006), who performed a series of analysis using the same data. This was because some variation was captured by SNP markers in the above estimation such that variance components accounted for by pedigree decreased, as expected. When markers were excluded from then model, $\hat{h}^2_{\text{ADD}}$ went up to 0.146, similar to the value reported in Valdar *et al.* (2006), and $\hat{h}^2_{\text{POE}}$ went up to 0.178.

Table 5.1: Average of testing set results of five 3-fold CV replicates in the mouse data (SE = standard error). ADD = additive model; POE = parent-of-origin effects model.

Model	$Corr^{(P)}_{\boldsymbol{y},\hat{\boldsymbol{y}}}$ (SE)[1]	$Corr^{(S)}_{\boldsymbol{y},\hat{\boldsymbol{y}}}$ (SE)[2]	MSE (SE)[3]
ADD	0.562 ($\pm$0.0040)	0.572 ($\pm$0.0036)	0.00244 ($\pm 1.65\times10^{-5}$)
POE	0.557 ($\pm$0.0060)	0.565 ($\pm$0.0060)	0.00246 ($\pm 2.45\times10^{-5}$)

[1] $Corr^{(P)}_{\boldsymbol{y},\hat{\boldsymbol{y}}}$: Pearson's correlation between observed and predicted value.
[2] $Corr^{(S)}_{\boldsymbol{y},\hat{\boldsymbol{y}}}$: Spearman's correlation between observed and predicted value.
[3] MSE: mean squared error.

Table 5.2: Estimated variance components ($\times 10^{-4}$) in the two models with all individuals included. ADD = additive model; POE = parent-of-origin effects model. NA = not applicable.

Model	$\hat{\sigma}_u^2$	$\hat{\sigma}_g^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_e^2$
ADD	2.60	NA	4.37	18.09
POE	2.13	1.05	4.26	16.74

5.4.2 Analysis of Simulated Data

In the simulation, five replicates were run, with each replicate coming from an independent run of QMSim simulation. Each of the five realizations had training and testing sample sizes of 2,000 individuals each; the number of SNPs was 10,000 and the number of QTLs in each replicate was 142, 167, 158, 141 and 149, respectively.

As described in Section 5.3.1, every replicate had 20 scenarios, each corresponding to a combination of ρ (imprinting level) and s (proportion of imprinted QTLs). When $\rho = 1$, however, three scenarios were actually redundant because in this case, all QTLs were unimprinted such that different values of s made no difference (Equation 5.7). Figure 5.1 displays the average prediction accuracy measured by Pearson's correlation between observed and predicted phenotypes in different simulation scenarios, and Figure 5.2 shows the MSE performance of the two models. Under both evaluation metrics, the ADD model performed better than the POE model when no imprinting was simulated ($\rho = 1$). When there was no parent-of-origin effects, the p extra parameters in the POE model led to overfitting of the training data, thus sacrificing predictive ability of future data. With parent-of-origin effects, the POE model outperformed the ADD model but dependent on the s and ρ settings. Typically, the POE model was better than the ADD model when ρ was small and s was large.

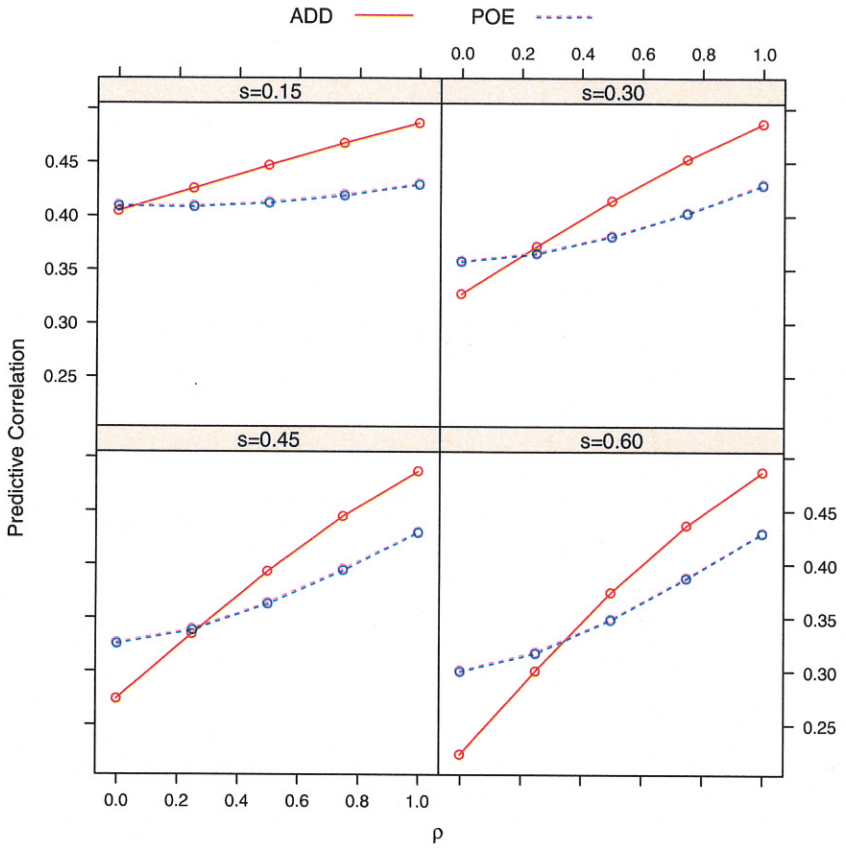


Figure 5.1: Average predictive correlation of two models measured by Pearson's correlation $(Corr^{(P)}_{\hat{\boldsymbol{y}},\boldsymbol{y}})$ between observed and predicted phenotype under different simulation settings. ADD = additive model; POE = parent-of-origin effects model. s = proportion of imprinted QTLs; $\rho = 0$ and $\rho = 1$ denote complete imprinting and no imprinting, respectively.

5.5 Discussion

5.5.1 Predictive Performance of the ADD and POE Models

Case 1: Complete imprinting ($\rho = 0$)

When imprinting was complete ($\rho = 0$), it was not surprising that the POE model performed better than the additive ADD model. The superiority of the POE model over the ADD model was dependent on the value of s: the larger the proportion of imprinted genes, the bigger the

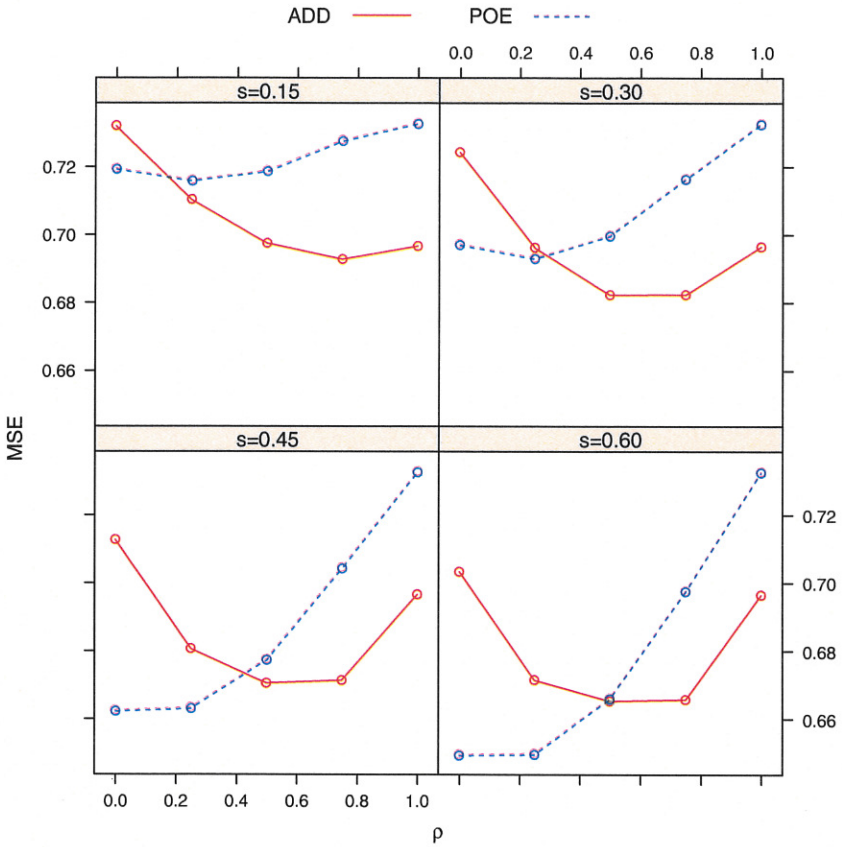


Figure 5.2: Averaged mean squared error (MSE) of two models between observed and predicted phenotype under different simulation settings. ADD = additive model; POE = parent-of-origin effects model. s = proportion of imprinted QTLs; $\rho = 0$ and $\rho = 1$ denote complete imprinting and no imprinting, respectively.

difference on predictive correlation and MSE between the two models. As s increased, a larger fraction of genetic variation was attributed to parent-of-origin effects, which cannot be captured by the ADD model. An interesting observation from Figures 5.1 and 5.2 is that for a given model, the predictive correlation and MSE decreased with an increase of s (Figure 5.3). Recall that the data was simulated such that the allelic effect was multiplied by ρ (less imprinting as $\rho \to 1$), and an s fraction of total QTLs were assumed to be imprinted (Equation 5.7). Suppose a single locus is maternally imprinted (the allele inherited from the mother written first), and that the values of the four genotypes (expressed as deviations from the population mean) are:

$$
\begin{aligned}
G_{11} &= \rho \cdot \theta_1 + \theta_1, \\
G_{21} &= \rho \cdot \theta_2 + \theta_1, \\
G_{12} &= \rho \cdot \theta_1 + \theta_2, \\
G_{22} &= \rho \cdot \theta_2 + \theta_2.
\end{aligned} \tag{5.8}
$$

Let p and q be the frequencies of the A_1 and A_2 alleles. The additive genetic variance at this locus can be calculated as:

$$
\begin{aligned}
\sigma_A^2 &= P_{11} \cdot G_{11}^2 + P_{21} \cdot G_{21}^2 + P_{12} \cdot G_{12}^2 + P_{22} \cdot G_{22}^2 \\
&= (1 + \rho^2) pq(\theta_1 - \theta_2)^2,
\end{aligned} \tag{5.9}
$$

where P_{ij} is the genotype frequency of $A_i A_j$ assuming Hardy-Weinberg equilibrium. Note that $\theta_1 - \theta_2$ is α, the allelic substitution effect defined by a standard additive genetic model. From Equation 5.9, when $\rho = 1$ (no imprinting), the expression yields $2pq\alpha^2$, the additive variance of a standard genetic model (e.g., Falconer and Mackay 1996; Lynch and Walsh 1998). When $\rho < 1$, however, this variance ("signal") is smaller than if not imprinted. Hence, for a given value of ρ that is smaller than 1 (0 in this case), the total variance of all QTL loci gets smaller as s increases. Since the environmental distribution was the same in all settings, heritability decreased with s, possibly producing a lower predictive ability.

Case 2: No imprinting ($\rho = 1$)

As stated above, when $\rho = 1$, the value of s does not affect the simulated data. In this simplest case, the ADD model outperformed the POE model in terms of predictive correlation and MSE, as the extra parameters in the POE model would capture noise only. In our Bayesian implementation of whole-genome prediction, genome-wide incorporation of parent-of-origin effects would result in an approximately doubled number of parameters relative to the ADD model. Therefore, it would

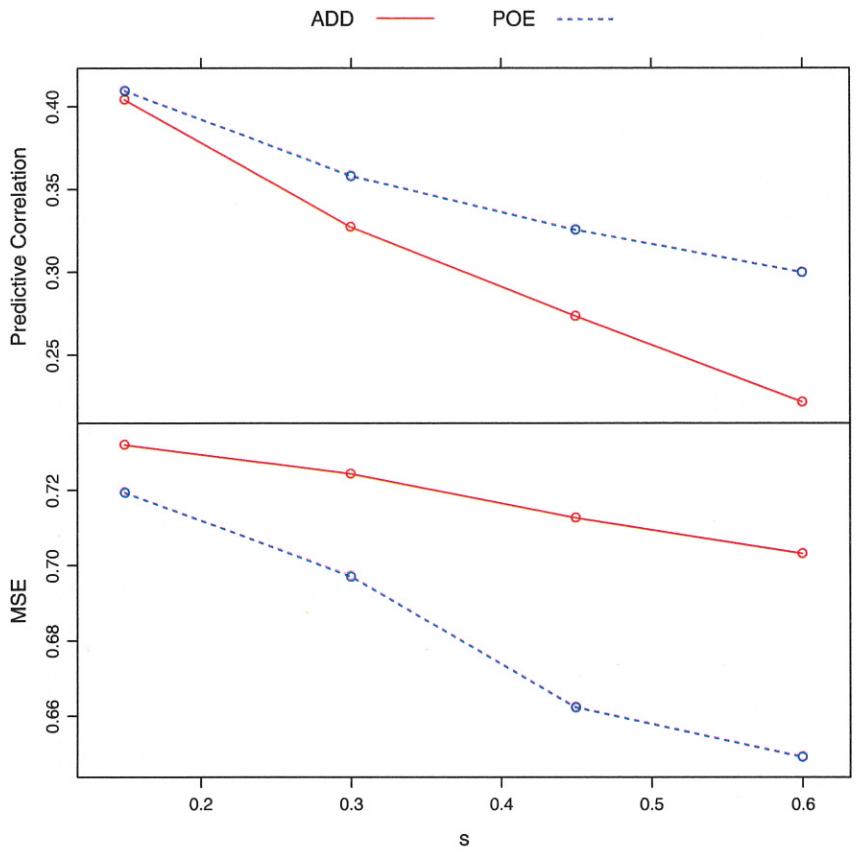


Figure 5.3: Trend of averaged predictive correlation and MSE with change of s (proportion of imprinted QTLs) under $\rho = 0$ (complete imprinting). Predictive correlation and MSE decrease as s goes up for both models. ADD = additive model; POE = parent-of-origin effects model.

be expected that the POE model produces overfitting. This can be seen in Figure 5.4: the training correlation of the POE model was always higher than that of the ADD model by about 4 percent, indicating a better fit to the training data. However, if, instead of capturing signal in the data, the better fit is due to higher model complexity, a penalty would be given to such model during the testing process (Hastie *et al.*, 2009), resulting in a lower predictive correlation ($\rho = 1$, Figure 5.1). Hence, the POE model was expected to have a higher prediction error than the ADD model when there were no parent-of-origin effects affecting the trait (Figure 5.2).

Overfitting might be a reason why the ADD model was better when the mouse data was used. Over replicates, the average training correlation using ADD and POE models was 0.830 ($\pm$0.014)

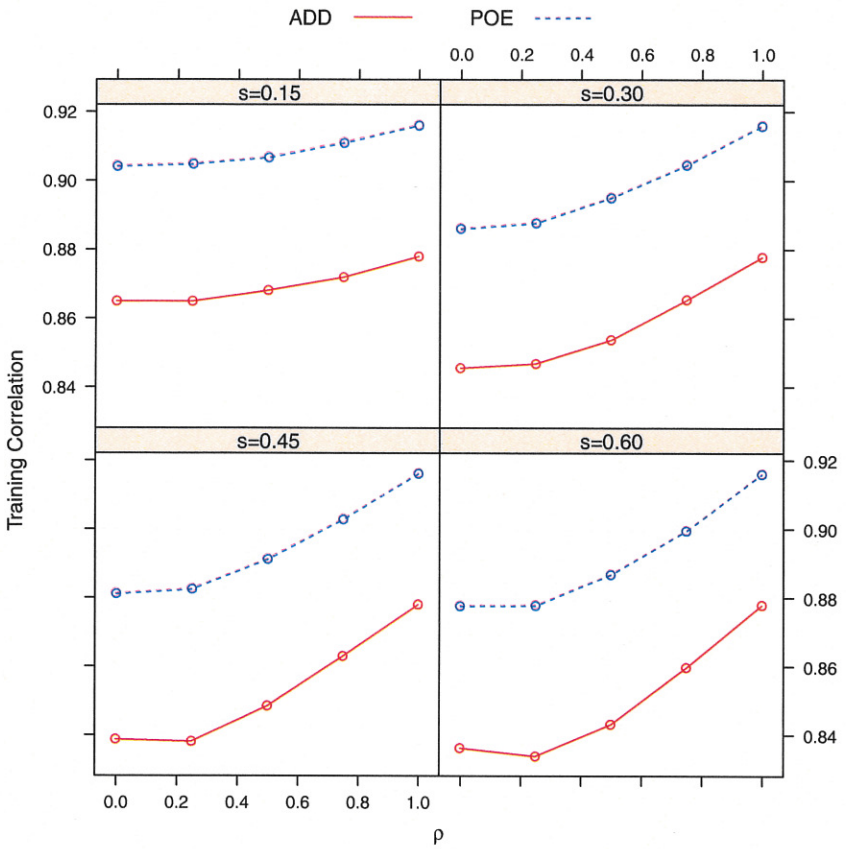


Figure 5.4: Training accuracy of two models measured by Pearson's correlation $\left(Corr_{\hat{\boldsymbol{y}},\boldsymbol{y}}^{(P)}\right)$ between observed and fitted phenotype under different simulation settings. ADD = additive model; POE = parent-of-origin effects model. s = proportion of imprinted QTLs; $\rho = 0$ and $\rho = 1$ denote complete imprinting and no imprinting, respectively.

and 0.858 (± 0.012), respectively, and the training MSE for the two models was 0.00127 ($\pm 9.1 \times 10^{-5}$) and 0.00113 ($\pm 8.2 \times 10^{-5}$), indicating that the POE model fitted the training data more closely. Technically, a more complex model would enhance prediction if true underlying signals are captured by the extra parameters, so that overfitting is not an issue. If, on the other hand, the true signal is not strong enough or training sample size is not large enough, overfitting would degrade prediction performance in the testing step. Although some parent-of-origin effects seem to exist, as indicated by the previous study using the same data (Chapter 4), these are not strong enough to overwhelm overfitting, resulting in a lower predictive performance.

Case 3: Partial imprinting

As imprinting changed from the highest ($\rho = 0$, complete imprinting) to the lowest level ($\rho = 1$, no imprinting), the predictive correlation of both models increased gradually for any given value of s, since total additive variance (signal) increased as ρ (Equation 5.9; left panel of Figure 5.5), so the predictive ability increased accordingly. Also, because the ADD model was better at $\rho = 1$ but the POE model was better at $\rho = 0$, curves representing the two models crossed at some point, and it was interesting to note that the cross point increased in terms of ρ (representing a less imprinting level) as s went up (Figure 5.1). Intuitively, the POE model would outperform the ADD model when the proportion of signal due to parent-of-origin effects reaches some threshold. Here, the variance accounted for by parent-of-origin effects is expressed as

$$\sigma_o^2 = \frac{1}{2}pq(\theta_1 - \theta_2)^2(1-\rho)^2 \tag{5.10}$$

according to the four genotypic values in Equation 5.8 and the one-locus imprinting model of Shete and Amos (2002), Spencer (2002), and de Koning *et al.* (2002); the ratio between Equations 5.10 and 5.9 gives the proportion of additive variance accounted for by parent-of-origin effect at that locus. For a larger s, this threshold is reached much faster than at a smaller s as $\rho \to 0$ (Figure 5.5, right panel), indicating that when fewer QTLs are imprinted, a higher imprinting level is needed for the POE model to gain advantage, as expected.

5.5.2 Proportion of Imprinted Genes

In our simulation, 0.15, 0.3, 0.45 and 0.6 were assigned to s (proportion of imprinted QTLs) in different scenarios. These values were chosen arbitrarily and are far beyond the proportion of imprinted genes with available evidence as, among approximately 25,000 human or murine genes, only about 200 have been identified as imprinted (`http://igc.otago.ac.nz/home.html`), i.e., 1% of the total number of genes. Even the smallest value of s chosen (0.15) is too large compared to

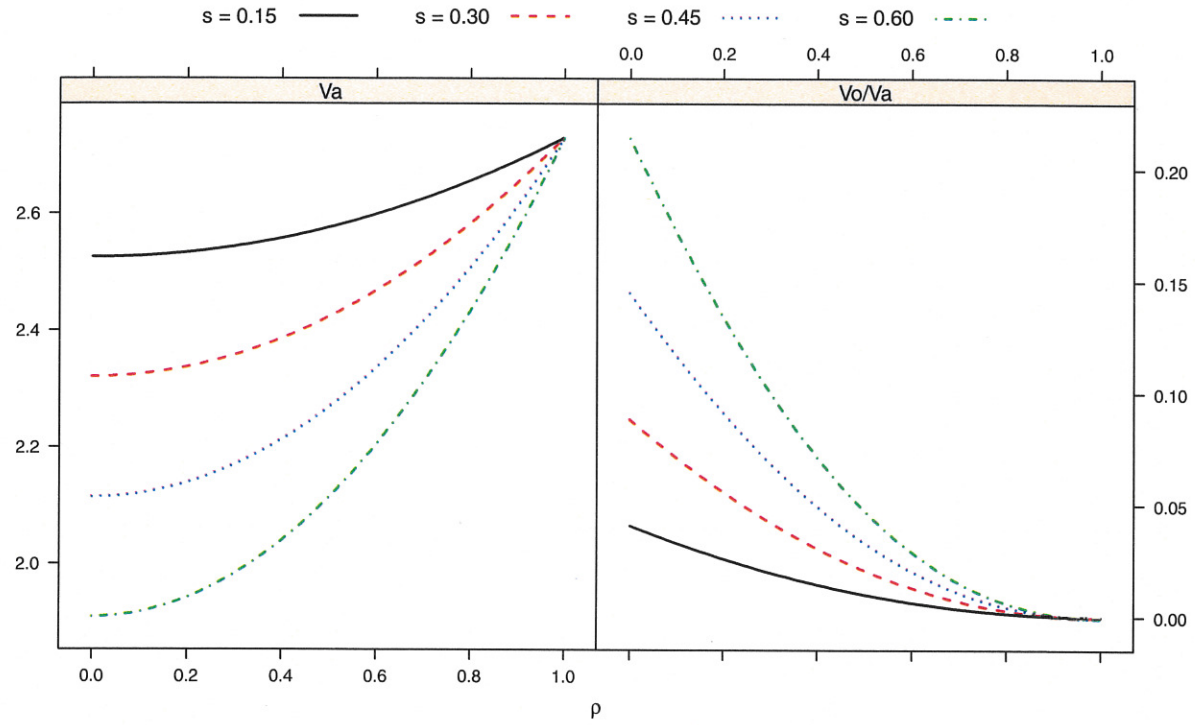


Figure 5.5: Stylized representation of the change of total additive variance across all 150 simulated QTL loci (V_a, left panel) and proportion of total additive variance due to parent-of-origin effects (V_o/V_a, right panel) at different values of ρ (imprinting level, changes from 0 to 1) and s (=$\{0.15, 0.3, 0.45, 0.6\}$, proportion of imprinted QTL).

this small fraction observed by far.

However, about 200 is the number of experimentally identified imprinted genes. This means that the function, expression profile and regulating mechanisms of such genes were assessed in well designed experiments, with verified imprinting status. It is possible that there are more imprinted genes in the mammalian genomes but not discovered so far. For example, Luedi *et al.* (2005) and Brideau *et al.* (2010) predicted that there might be hundreds of imprinted genes in the murine genome. Unfortunately, no consensus estimate on the number of imprinted genes in the mammalian genome is available (Kelsey and Bartolomei, 2012). Further, imprinting might be more prevalent than previously thought, as argued in several review studies (e.g., Sha, 2008; Lawson *et al.*, 2013). Specifically, in a total of 127 detected metabolic-related QTLs, about 60% had imprinting effects. In an earlier study, 54% of 602 genes expressed in human kidney or liver tissues were shown to have

strong parent-of-origin effects caused by preferential expression, with some of them not located in known imprinted genomic regions (Lo *et al.*, 2003). Given these studies, therefore, we decided to increase the proportion of imprinted QTLs in our simulation over the 1% mentioned earlier.

Further, for the approximately 200 identified imprinted genes, a vast majority are growth- and/or development-related. This was noticed when the famous "parent-offspring conflict hypothesis" was proposed (Haig and Westoby, 1989; Haig and Westoby, 1991; Moore and Haig, 1991) to explain the evolution of imprinting. Although Lush often stated the view that all complex traits are possibly affected by all genes at various degrees (Lush, 1945, 1948), it is unlikely that all tens of thousands of genes in the mammalian genome affect a trait jointly (Gianola and Rosa, 2015). Since there is no consensus on how many genes affect specific complex traits, tens to several hundreds might be a reasonable guess. Hence, within the hundreds of genes that control a single trait, say, fetal growth, it is possible that a considerable proportion is subject to imprinting. In addition, imprinting is a major cause of parent-of-origin effects, but not the only one (Guilmatre and Sharp, 2012). Therefore, when imprinting was considered as the only cause to simplify the source of parent-of-origin effects in the simulation, we set the proportion of imprinted QTLs up to 60 percent.

5.5.3 Information other than DNA Polymorphisms

Incorporating parent-of-origin effects into a prediction model may be helpful if it accounts for a considerable proportion of the total variance. In practice, additive variance is the major contributor to phenotypic variability for most complex traits (Hill *et al.*, 2008). Along with the overfitting problems associated with the POE model, the preceding implies that the POE model may bring only a minimal advantage in most cases. Therefore, it might be helpful to consider other sources of information in whole genome prediction to incorporate parent-of-origin effects. Since epigenetics, a main cause of parent-of-origin effects, is the study of heritable variation that does not involve a change of DNA sequence (Riggs *et al.*, 1996; Riggs and Porter, 1996; Naumova and Greenwood,

2013), our prediction model may fail under many situations because only variation at the DNA level (e.g., SNP markers) is used as input. Hence, incorporating epigenetic information as a supplement to SNP markers might be useful (González-Recio, 2012). An example is the success of epigenome-wide association studies (EWAS) that use epigenetic biomarkers to find disease-related genomic regions (Rakyan *et al.*, 2011; Bell, 2013; Flanagan, 2015).

Including epigenetic information in whole-genome prediction is promising but might also be challenging. One aspect is the amount of information one needs to deal with. Consider DNA methylation as an example: it is the addition of a methyl group to either the 5-position carbon atom of the cytosine pyrimidine ring, or to the 6-position nitrogen atom of the adenine purine ring, with the latter observed mainly in mitochondrial DNA of flowering plants (Vanyushin, 2006). Two important features of DNA methylation are: 1) it is tissue and developmental-stage specific; 2) it is reversible as the added methyl group can be removed from the methylated DNA molecule. Due to this second feature, methylation status is unstable compared to DNA polymorphisms and, for a certain cytosine locus, it may shift between methylated and unmethylated states. Thus, even though modern technologies are able to convert the unstable methylation information into stable sequence information via bisulfite treatment (Frommer *et al.*, 1992; Bock, 2012), the methylation profile is for a specific time in a specific sample of cells. The term "methylome" is thus misused: in many studies, it is actually referring to a "snap shot" of the entire methylome at a certain time point from a certain tissue given the first feature of DNA methylation. Comparing to DNA sequence information which is size-invariant (unless a somatic mutation occurs) throughout an individual's life time, the size of the methylome is highly variable and can be extremely large. Along with other epigenetic mechanisms like histone modification, the size of the human epigenome is potentially enormous. For example, the diploid human epigenome contains more than 10^8 cytosins (of which $> 10^7$ are found in CpG dinucleotides, the major target of mammalian DNA methylation) and more than 10^8 histone tails (the target of histone modification) that can all potentially vary (Rakyan *et al.*, 2011). It has been estimated that the human epigenome could be thousands of times larger

than the genome (Zhang and Jeltsch, 2010)! Given this magnitude, choosing appropriate epigenetic information from a suitable tissue is crucial, and powerful and reliable analytical tools must be developed to assure an appropriate use of the information.

Apart from the size of the epigenome, epigenetic mechanisms are affected by environmental effects. For instance, the methyl group added to a DNA molecule must come from a methyl group donor. One major source is the diet (Niculescu and Zeisel, 2002), so different diets can result in different methylation profiles that lead to different phenotypes. Several cases demonstrate the the impact of nutrition on epigenetics. In mice, the coat color of genetically identical individuals showed variation when their mothers were fed differently during pregnancy (Morgan *et al.*, 1999; Dolinoy *et al.*, 2006). In honey bees, almost all female individuals in a colony are (almost, if not exactly) genetically identical. However, the royalactin found in royal jelly turns one (and only one) individual into a queen and the rest remain as workers (Kamakura, 2011). In livestock, maternal diet during pregnancy can alter the DNA methylation of the fetus and, hence, result in changes in gene expression (Lan *et al.*, 2013). This evidence indicates that environmental variation brings extra difficulties to the already complicated epigenetic analysis.

Furthermore, epigenome profiling is very expensive. In the case of methylation, due to the massive number of CpG sites within the mammalian genome, high resolution methylation profiles are very costly. Although reduced representation bisulfite sequencing (RRBS, Meissner *et al.*, 2005) can reduce the profiling costs by selecting a small proportion of representative CpG sites from certain regions (e.g., gene promoter regions) of the genome, methylation profiling of a large cohort (e.g., thousands of individuals in a WGP study) is still expensive, especially when multiple "snapshots" of the methylome are to be considered.

In short, epigenetic polymorphisms could contribute to genetic studies and open a door to a better understanding of biological systems. However, many challenges need to be resolved before this information can be efficiently used to advantage.

5.6 Conclusion

We proposed a model that is capable of incorporating parent-of-origin effects into whole genome prediction using pedigree and DNA information. The model is based on Bayesian regression implemented via MCMC. The POE model was compared with an additive model using real and simulated data. Results from the real data analysis suggested that the POE model did not outperform the ADD model, because the proportion of genetic variance explained by parent-of-origin effects was not large enough. Also, a doubled number of parameters in the POE model possibly produced overfitting in the model training step, resulting in a reduced prediction ability. However, incorporating parent-of-origin effects in whole genome prediction led to an increased heritability estimate, which was consistent with other studies (e.g., Neugebauer *et al.*, 2010a,b; Mott *et al.*, 2014).

Simulated data was used to evaluate model performance under different conditions. The data was simulated using a one-locus imprinting model (extended to several loci) to simplify the source of parent-of-origin effects. The model was evaluated under different proportions of imprinted loci and various imprinting levels. When parent-of-origin effects contributed a large proportion of genetic variation, the POE model performed better than the ADD model, as expected. However, the POE model was not better than the ADD model when no parent-of-origin effects were contributing to the trait, probably due to overfitting.

Owing to the discovery of more imprinted genes and of parent-of-origin-effects-affected complex traits, obtaining predictions with consideration of parent-of-origin effects was deemed attractive. However, results from our simulation suggested that it did not always work well unless parent-of-origin effects contributed to the complex trait sizably. Thus, assessing parent-of-origin effects based on additional information to DNA polymorphisms might be helpful. Because many technical challenges need to be faced at the current stage of knowledge, future studies need to explore more effective prediction machines for parent-of-origin-effects-affected complex traits in animals, plants, and humans.

References

Abramowitz, L. K. and M. S. Bartolomei, 2009. An in vitro ES cell imprinting model shows that imprinted expression of the *Igf2r* gene arises from an allele-specific expression bias. Development, 136: 437–448

Barlow, D. P., 2011. Genomic Imprinting: A Mammalian Epigenetic Discovery Model. Annu. Rev. Genet., 45: 379–403

Barlow, D. P. and M. S. Bartolomei, 2014. Genomic imprinting in mammals. Cold Spring Harb. Perspect. Biol., 6(2): a018382

Bell, C. G., 2013. Epigenome-Wide Association Studies: Potential Insights into Human Disease. In A. Naumova and C. Greenwood (Editors), *Epigenetic and Complex Traits*, pp. 287–317. Springer New York

Bock, C., 2012. Analysing and interpreting DNA methylation data. Nat. Rev. Genet., 13(10): 705–719

Brideau, C. M., K. E. Eilertson, J. A. Hagarman, *et al.*, 2010. Successful computational prediction of novel imprinted genes from epigenomic features. Mol. Cell Biol., 30(13): 3357–3370

Cheverud, J. M., R. Hager, C. Roseman, *et al.*, 2008. Genomic imprinting effects on adult body composition in mice. Proc. Natl. Acad. Sci. U.S.A., 105(11): 4253–4258

Clayton-Smith, J., 2003. Genomic imprinting as a cause of disease. BMJ, 327(7424): 1121–1122

Coster, A., O. Madsen, H. C. Heuven, *et al.*, 2012. The imprinted gene *DIO3* is a candidate gene for litter size in pigs. PLoS ONE, 7(2): e31825

Cui, Y., J. M. Cheverud, and R. Wu, 2007. A statistical model for dissecting genomic imprinting through genetic mapping. Genetica, 130(3): 227–239

Cui, Y., Q. Lu, J. M. Cheverud, *et al.*, 2006. Model for mapping imprinted quantitative trait loci in an inbred F_2 design. Genomics, 87(4): 543–551

de Koning, D. J., H. Bovenhuis, and J. A. van Arendonk, 2002. On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. Genetics, 161(2): 931–938

de los Campos, G., D. Gianola, and G. J. Rosa, 2009a. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J. Anim. Sci., 87(6): 1883–1887

de los Campos, G., J. M. Hickey, R. Pong-Wong, *et al.*, 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193(2): 327–345

de los Campos, G., H. Naya, D. Gianola, *et al.*, 2009b. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182(1): 375–385

de los Campos, G. and P. Pérez Rodríguez, 2014. *BGLR: Bayesian Generalized Linear Regression.*

R package version 1.0.3

Dekkers, J. C., 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J. Anim. Sci., 82 (e-Suppl.): 313–328

Delaval, K. and R. Feil, 2004. Epigenetic regulation of mammalian genomic imprinting. Curr. Opin. Genet. Dev., 14(2): 188–195

Dolinoy, D. C., J. R. Weidman, R. A. Waterland, *et al.*, 2006. Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. Environ. Health Perspect., 114(4): 567–572

Engellandt, T. H. and B. Tier, 2002. Genetic variances due to imprinted genes in cattle. J. Anim. Breed. Genet., 119(3): 154–165

Erbe, M., E. C. G. Pimentel, A. R. Sharifi, *et al.*, 2010. Assessment of cross-validation strategies for genomic prediction in cattle. In *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*. Leipzig, Germany

Essl, A. and K. Voith, 2002. Genomic imprinting effects on dairy- and fitness-related traits in cattle. J. Anim. Breed. Genet., 119(3): 182–189

Falconer, D. S. and T. F. C. Mackay, 1996. *Introduction to Quantitative Genetics*. Prentice Hall, 4th edition

Falls, J. G., D. J. Pulford, A. A. Wylie, *et al.*, 1999. Genomic imprinting: implications for human disease. Am. J. Pathol., 154(3): 635–647

Flanagan, J. M., 2015. Epigenome-wide association studies (EWAS): past, present, and future. In M. Verma (Editor), *Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment, and Prognosis*, Methods in Molecular Biology (vol. 1238), pp. 51–63. Humana Press

Frésard, L., M. Morisson, J. M. Brun, *et al.*, 2013. Epigenetics and phenotypic variability: some interesting insights from birds. Genet. Sel. Evol., 45: 16

Frommer, M., L. E. McDonald, D. S. Millar, *et al.*, 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. U.S.A., 89(5): 1827–1831

Gianola, D., 2013. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics, 194(3): 573–596

Gianola, D. and G. de los Campos, 2008. Inferring genetic values for quantitative traits non-parametrically. Genet. Res. (Camb), 90(6): 525–540

Gianola, D., G. de los Campos, W. G. Hill, *et al.*, 2009. Additive genetic variability and the Bayesian alphabet. Genetics, 183(1): 347–363

Gianola, D., R. L. Fernando, and A. Stella, 2006. Genomic-assisted prediction of genetic value with

semiparametric procedures. Genetics, 173(3): 1761–1776

Gianola, D., H. Okut, K. A. Weigel, *et al.*, 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet., 12: 87

Gianola, D. and G. J. M. Rosa, 2015. One hundred years of statistical developments in animal breeding. Annu. Rev. Anim. Biosci., 3: 19–56

Gibson, J., B. Kennedy, L. Schaeffer, *et al.*, 1988. Gametic models for estimation of autosomally inherited genetic effects that are expressed only when received from either a male or female parent. J. Dairy Sci., 71 (Suppl. 1): 143 (Abstr.)

Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica, 136(2): 245–257

González-Camacho, J. M., G. de Los Campos, P. Pérez, *et al.*, 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet., 125(4): 759–771

González-Recio, O., 2012. Epigenetics: a new challenge in the post-genomic era of livestock. Front. Genet., 2: 106

Guilmatre, A. and A. J. Sharp, 2012. Parent of origin effects. Clin. Genet., 81(3): 201–209

Guimarães, E., J. Ruane, B. Scherf, *et al.* (Editors), 2007. *Marker-assisted Selection: Current Status and Future Perspectives in Crops, Livestock, Forestry and Fish.* Food and Agriculture Organization of the United Nations

Hager, R., J. M. Cheverud, and J. B. Wolf, 2008. Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. Genetics, 178(3): 1755–1762

Haig, D. and M. Westoby, 1989. Parent-specific gene expression and the triploid endosperm. Am. Nat., 134(1): 147–155

Haig, D. and M. Westoby, 1991. Genomic imprinting in endosperm: its effect on seed development in crosses between species, and between different ploidies of the same species, and its implications for the evolution of apomixis. Philos. Trans. R. Soc. Lond., B, Biol. Sci., 333: 1–13

Hall, J. G., 1990. Genomic imprinting: review and relevance to human diseases. Am. J. Hum. Genet., 46(5): 857–873

Hastie, T., R. Tibshirani, and J. Friedman, 2009. *The elements of statistical learning: data mining, inference and prediction.* Springer, 2nd edition

Henderson, C. R., 1984. *Applications of Linear Models in Animal Breeding.* University of Guelph, Ontario, Canada

Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet., 4(2): e1000008

Holl, J. W., J. P. Cassady, D. Pomp, *et al.*, 2004. A genome scan for quantitative trait loci and

imprinted regions affecting reproduction in pigs. J. Anim. Sci., 82(12): 3421–3429

Imumorin, I. G., E. H. Kim, Y. M. Lee, *et al.*, 2011. Genome scan for parent-of-origin QTL effects on bovine growth and carcass traits. Front. Genet., 2: 44

Jeon, J. T., O. Carlborg, A. Tornsten, *et al.*, 1999. A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the *IGF2* locus. Nat. Genet., 21(2): 157–158

Jonas, E. and D. J. de Koning, 2013. Does genomic selection have a future in plant breeding? Trends Biotechnol., 31(9): 497–504

Kamakura, M., 2011. Royalactin induces queen differentiation in honeybees. Nature, 473(7348): 478–483

Kärst, S., A. R. Vahdati, G. A. Brockmann, *et al.*, 2012. Genomic imprinting and genetic effects on muscle traits in mice. BMC Genomics, 13: 408

Kelsey, G. and M. S. Bartolomei, 2012. Imprinted genes ... and the number is? PLoS Genet., 8(3): e1002601

Khatib, H., 2007. Is it genomic imprinting or preferential expression? Bioessays, 29(10): 1022–1028

Kilpinen, H. and E. T. Dermitzakis, 2012. Genetic and epigenetic contribution to complex traits. Hum. Mol. Genet., 21(R1): R24–28

Kim, E. H., B. H. Choi, K. S. Kim, *et al.*, 2007. Detection of mendelian and parent-of-origin quantitative trait loci in a cross between korean native pig and landrace. I. Growth and body composition traits. Asian-Aust. J. Anim. Sci., 20: 669–676

Knott, S. A., L. Marklund, C. S. Haley, *et al.*, 1998. Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. Genetics, 149(2): 1069–1080

Lan, X., E. C. Cretney, J. Kropp, *et al.*, 2013. Maternal diet during pregnancy induces gene expression and DNA methylation changes in fetal tissues in sheep. Front. Genet., 4: 49

Lawson, H. A., J. M. Cheverud, and J. B. Wolf, 2013. Genomic imprinting and parent-of-origin effects on complex traits. Nat. Rev. Genet., 14(9): 609–617

Lee, H. K., S. S. Lee, T. H. Kim, *et al.*, 2003. Detection of imprinted quantitative trait loci (QTL) for growth traits in pigs. Asian-Aust. J. Anim. Sci., 16: 1087–1092

Legarra, A., I. Aguilar, and I. Misztal, 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci., 92(9): 4656–4663

Lewis, A. and L. Redrup, 2005. Genetic imprinting: conflict at the Callipyge locus. Curr. Biol., 15(8): R291–294

Li, E., C. Beard, and R. Jaenisch, 1993. Role for DNA methylation in genomic imprinting. Nature, 366(6453): 362–365

Liu, T., R. J. Todhunter, S. Wu, *et al.*, 2007. A random model for mapping imprinted quantitative trait loci in a structured pedigree: an implication for mapping canine hip dysplasia. Genomics, 90(2): 276–284

Lo, H. S., Z. Wang, Y. Hu, *et al.*, 2003. Allelic variation in gene expression is common in the human genome. Genome Res., 13(8): 1855–1862

Luedi, P. P., A. J. Hartemink, and R. L. Jirtle, 2005. Genome-wide prediction of imprinted murine genes. Genome Res., 15(6): 875–884

Lush, J. L., 1945. *Animal Breeding Plans.* Iowa State College Press, Ames, IA, 3rd edition

Lush, J. L., 1948. *The Genetics of Populations* (Mimeo). Iowa State Unviersity, Ames, IA

Lynch, M. and B. Walsh, 1998. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates

McEwen, K. R. and A. C. Ferguson-Smith, 2009. Genomic imprinting – a model for roles of histone modifications in epigenetic control. In A. C. Ferguson-Smith, J. M. Greally, and R. A. Martienssen (Editors), *Epigenomics*, pp. 235–258. Springer Netherlands

Meijers-Heijboer, E. J., L. A. Sandkuijl, H. G. Brunner, *et al.*, 1992. Linkage analysis with chromosome 15q11-13 markers shows genomic imprinting in familial Angelman syndrome. J. Med. Genet., 29(12): 853–857

Meissner, A., A. Gnirke, G. W. Bell, *et al.*, 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res., 33(18): 5868–5877

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4): 1819–1829

Meyer, K. and B. Tier, 2012. Estimates of variances due to parent of origin effects for weights of Australian beef cattle. Anim. Prod. Sci., 52: 215–224

Moore, T. and D. Haig, 1991. Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet., 7(2): 45–49

Morcos, L., B. Ge, V. Koka, *et al.*, 2011. Genome-wide assessment of imprinted expression in human cells. Genome Biol., 12(3): R25

Morgan, H. D., H. G. Sutherland, D. I. Martin, *et al.*, 1999. Epigenetic inheritance at the agouti locus in the mouse. Nat. Genet., 23(3): 314–318

Morota, G. and D. Gianola, 2014. Kernel-based whole-genome prediction of complex traits: a review. Front. Genet., 5: 363

Mott, R., W. Yuan, P. Kaisaki, *et al.*, 2014. The architecture of parent-of-origin effects in mice. Cell, 156(1-2): 332–342

Mrode, R., 2014. *Linear Models for the Prediction of Animal Breeding Values.* CAB International, 3rd edition

Nakaya, A. and S. N. Isobe, 2012. Will genomic selection be a practical method for plant breeding? Ann. Bot., 110(6): 1303–1316

Naumova, A. and C. Greenwood (Editors), 2013. *Epigenetics and Complex Traits.* Springer New York. ISBN 9781461480785

Neugebauer, N., H. Luther, and N. Reinsch, 2010a. Parent-of-origin effects cause genetic variation in pig performance traits. Animal, 4(5): 672–681

Neugebauer, N., I. Räder, H. J. Schild, *et al.*, 2010b. Evidence for parent-of-origin effects on genetic variability of beef traits. J. Anim. Sci., 88(2): 523–532

Nezer, C., L. Moreau, B. Brouwers, *et al.*, 1999. An imprinted QTL with major effect on muscle mass and fat deposition maps to the *IGF2* locus in pigs. Nat. Genet., 21(2): 155–156

Nicholls, R. D., S. Saitoh, and B. Horsthemke, 1998. Imprinting in Prader-Willi and Angelman syndromes. Trends Genet., 14(5): 194–200

Niculescu, M. D. and S. H. Zeisel, 2002. Diet, methyl donors and DNA methylation: interactions between dietary folate, methionine and choline. J. Nutr., 132 (Suppl. 8): 2333S–2335S

Nolan, C. M., J. K. Killian, J. N. Petitte, *et al.*, 2001. Imprint status of *M6P/IGF2R* and *IGF2* in chickens. Dev. Genes Evol., 211(4): 179–183

O'Neill, M. J., R. S. Ingram, P. B. Vrana, *et al.*, 2000. Allelic expression of *IGF2* in marsupials and birds. Dev. Genes Evol., 210(1): 18–20

Park, T. and G. Casella, 2008. The Bayesian Lasso. J. Amer. Statist. Assoc., 103(482): 681–686

Pérez, P. and G. de los Campos, 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics, 198(2): 483–495

Pérez-Rodríguez, P., D. Gianola, J. M. González-Camacho, *et al.*, 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 (Bethesda), 2(12): 1595–1605

Rakyan, V. K., T. A. Down, D. J. Balding, *et al.*, 2011. Epigenome-wide association studies for common human diseases. Nat. Rev. Genet., 12(8): 529–541

Reik, W. and J. Walter, 2001. Genomic imprinting: parental influence on the genome. Nat. Rev. Genet., 2(1): 21–32

Ribaut, J.-M. and D. A. Hoisington, 1998. Marker-assisted selection: new tools and strategies. Trends Plant Sci., 3: 236–239

Riggs, A. D., R. A. Martienssen, and V. E. A. Russo, 1996. Introduction. In V. E. A. Russo, R. A. Martienssen, and A. D. Riggs (Editors), *Epigenetic Mechanisms of Gene Regulation*, pp. 1–4. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Riggs, A. D. and T. N. Porter, 1996. Overview of epigenetic mechanisms. In V. E. A. Russo, R. A.

Martienssen, and A. D. Riggs (Editors), *Epigenetic Mechanisms of Gene Regulation*, pp. 29–45. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Sargolzaei, M. and F. S. Schenkel, 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25(5): 680–681

Schaeffer, L., B. Kennedy, and J. Gibson, 1989. The inverse of the gametic relationship matrix. J. Dairy Sci., 72(5): 1266–1272

Schaeffer, L. R., 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet., 123(4): 218–223

Sha, K., 2008. A mechanistic view of genomic imprinting. Annu. Rev. Genomics Hum. Genet., 9: 197–216

Shete, S. and C. I. Amos, 2002. Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. Am. J. Hum. Genet., 70(3): 751–757

Solter, D., 1992. Relevance of genomic imprinting to human diseases. Curr. Opin. Biotechnol., 3(6): 632–636

Spencer, H. G., 2002. The correlation between relatives on the supposition of genomic imprinting. Genetics, 161(1): 411–417

Stella, A., K. J. Stalder, A. M. Saxton, *et al.*, 2003. Estimation of variances for gametic effects on litter size in Yorkshire and Landrace swine. J. Anim. Sci., 81(9): 2171–2178

Thomsen, H., H. K. Lee, M. F. Rothschild, *et al.*, 2004. Characterization of quantitative trait loci for growth and meat quality in a cross between commercial breeds of swine. J. Anim. Sci., 82(8): 2213–2228

Tuiskula-Haavisto, M., D. J. de Koning, M. Honkatukia, *et al.*, 2004. Quantitative trait loci with parent-of-origin effects in chicken. Genet. Res., 84(1): 57–66

Tuiskula-Haavisto, M. and J. Vilkki, 2007. Parent-of-origin specific QTL – a possibility towards understanding reciprocal effects in chicken and the origin of imprinting. Cytogenet. Genome Res., 117(1-4): 305–312

Úbeda, F. and J. F. Wilkins, 2008. Imprinted genes and human disease: an evolutionary perspective. In J. F. Wilkins (Editor), *Genomic Imprinting*, volume 626 of *Advances in experimental medicine and biology*, pp. 101–115. Springer, New York & Landes Bioscience, Austin, TX

Valdar, W., L. C. Solberg, D. Gauguier, *et al.*, 2006. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat. Genet., 38(8): 879–887

VanRaden, P. M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci., 91(11): 4414–4423

Vanyushin, B. F., 2006. DNA methylation in plants. Curr. Top. Microbiol. Immunol., 301: 67–122

Wimmer, V., T. Albrecht, H.-J. Auinger, *et al.*, 2012. synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics, 28(15): 2086–2087

Wolf, J. B., J. M. Cheverud, C. Roseman, *et al.*, 2008a. Genome-wide analysis reveals a complex pattern of genomic imprinting in mice. PLoS Genet., 4(6): e1000091

Wolf, J. B., R. Hager, and J. M. Cheverud, 2008b. Genomic imprinting effects on complex traits: a phenotype-based perspective. Epigenetics, 3(6): 295–299

Young, N. D., 1999. A cautiously optimistic vision for marker-assisted breeding. Mol. Breeding, 5(6): 505–510

Zhang, Y. and A. Jeltsch, 2010. The application of next generation sequencing in dna methylation analysis. Genes, 1(1): 85–101

Chapter 6

Non-parametric Prediction of Plant Height in *Arabidopsis thaliana* Using DNA Methylation Data

Prediction of complex traits using molecular genetic information is an active area in quantitative genetics research. In the post-genomic era, many types of -omic data (e.g., transcriptomic, epigenomic, methylomic, proteomic) are becoming increasingly available. Therefore, evaluating the utility of this massive amount of information in prediction of complex traits is of interest. DNA methylation, the covalent change of a DNA molecule without affecting its underlying sequence, is one quantifiable form of epigenetic modification. In this chapter, methylation data was used to perform a non-parametric prediction of plant height in *Arabidopsis thaliana* using reproducing kernel Hilbert spaces (RKHS) regression.

6.1 Introduction

Epigenetics focuses on heritable changes of genetic materials that do not reside in the sequence of DNA, called epigenetic modifications (Riggs *et al.*, 1996; Riggs and Porter, 1996). Major forms of these changes are DNA methylation, histone modification and non-coding RNAs (ncRNAs) (Rivera and Bennett, 2010). DNA methylation is the most common epigenetic modification, which can have various forms depending on the targeting nucleotide of the modification (Ratel *et al.*, 2006). In vertebrates and flowering plants, it is usually referred to as the covalent addition of a methyl group (-CH_3) to the 5-position carbon atom (^{5}C) of the cytosine pyrimidine ring, resulting in 5-methylcytosine (^{m5}C) (Jeltsch, 2002; Meissner *et al.*, 2005; Vanyushin, 2006). Thus, "DNA methylation" stands for ^{m5}C throughout this chapter. Histone modification is the multivalent modification of histone tails of the core histones, which can be acetylation, methylation, phosphorylation, ubiquitination, and symoylation (Kouzarides, 2007; Ruthenburg *et al.*, 2007). Both DNA methylation and histone modification interact with the entering and binding of transcription factors (TF) to the DNA molecule such that gene expression is altered. Usually, DNA methylation is associated with reduced gene expression (Bird, 1984; Razin and Cedar, 1991; Lim and Maher, 2010) and histone modification can either enhance or repress expression according to different modification targets (e.g., which amino acid are at the histone tail) and modification types (e.g., methylation or acetylation) (Berger, 2002; Cheung and Lau, 2005). Recently, ncRNAs were found to comprise a hidden layer of internal signals that control various levels of gene expression associated with physiological and developmental processes. Their role in epigenetic regulation of gene expression has been acknowledged as well (Zhou *et al.*, 2010; Kaikkonen *et al.*, 2011).

Characteristics of epigenetic modifications determine their important role on gene expression regulation, where they play as an "operating system" that controls the functioning of the DNA and genes. Thus, malfunctioning of the regulation process can have severe consequences. In epidemiology and human genetics, many diseases and disorders, including cancer, have been confirmed to

have an epigenetic basis (Jones and Baylin, 2002, 2007; Jiang *et al.*, 2004; Esteller, 2008; Tollefsbol, 2012; Pembrey, 2012). For example, the Prader-Willi (PWS) and Angelman (AS) syndromes are sister imprinting-related disorders involving deletion of DNA segments derived from different parent at the same genomic region (Meijers-Heijboer *et al.*, 1992; Nicholls *et al.*, 1998; Cassidy *et al.*, 2000). Another good example of epigenetics-related diseases is that of oncogenes, which exist in almost everyone's genome, but only a small proportion of the population develops a cancer. This is because the promoter region of a tumor suppressor gene is usually unmethylated such that the gene is expressed normally and, therefore, it prohibits the formation of a tumor. In cases where there is a hyper-methylation in the promoter region, the tumor suppressor is deactivated and a cancer develops (Robertson, 2002; Jones and Baylin, 2002; Egger *et al.*, 2004). Due to the potentially important role of epigenetics in diseases, epigenome-wide association studies (EWAS), a counterpart of genome-wide association studies (GWAS) at the epigenome level, has been conducted in recent years (MacArthur, 2008; Rakyan *et al.*, 2011; Bell, 2013), aiming at finding associations between epigenetic polymorphisms and traits of interest, instead of using DNA polymorphisms (e.g., SNPs). Although epigenetic regulation is not restricted to DNA methylation, the latter is the most commonly used biomarker in EWAS at present, because it is more stable and easier to be quantified than other epigenetic regulatory mechanisms (Flanagan, 2015). In EWAS, DNA methylation across the whole genome is converted into a certain measurement reflecting the "methylation level" using either methylation-sensitive enzyme digestion (Waalwijk and Flavell, 1978; Kaput and Sneider, 1979), methylated DNA immunoprecipitation (MeDIP) (Weber *et al.*, 2005) or bisulfite sequencing (BS-Seq) that combines next-generation sequencing techniques with bisulfite conversion (Frommer *et al.*, 1992), with BS-Seq being the most popular method used in methylation profiling. In BS-Seq, a DNA sample is treated with sodium bisulfite, which can convert unmethylated cytosine into uracil whereas methylated cytosine is intact. Uracil is read as thymine in PCR (polymerase chain reaction) and sequence alignment after PCR amplification gives the counts of C (originally methylated cytosine) and T (originally unmethylated cytosine) at a single-base resolution. The ratio $\frac{C}{C+T}$ gives the absolute methylation level at that base, which is referred to as the β value in

methylation profiling literature, and is usually considered as the "gold standard" in methylation quantification (Krueger *et al.*, 2012). Once the methylation level is obtained, statistical methods are then applied to find associations between the "methylation profile" and the trait of interest in a selected sample.

Although some diseases associated with the dysregulation of epigenetic modification at some genomic region have been found, EWAS has the same drawbacks as GWAS: it is difficult to estimate how much variation in the phenotypes, especially for complex traits, is explained by epigenetic polymorphisms, even if there is evidence that it contributes to the phenotype, either biologically or statistically. Two studies addressing this question have been published in recent years, with *Arabidopsis thaliana* used as experimental material (Johannes *et al.*, 2009; Reinders *et al.*, 2009). In Johannes *et al.* (2009), a wild-type inbred line was chosen as the paternal founder and a *ddm1* mutant was used as the maternal founder. The *ddm1* mutant was genetically identical to the wild-type, except for the *DDM1* locus and a few other loci. The *DDM1* locus encodes an ATPase chromatin remodeler that is involved in methylation maintenance; and the *ddm1* mutant used in their study was featured by a whole-genome-wide demethylation. The F_1 generation was obtained by crossing the wild-type (as male) and *ddm1* mutant (as female), and then it was backcrossed with the wild type (as male) to create the F_2 generation. The F_2 individuals were selfed for several generations to construct a population of epigenetic recombinant inbred lines (epiRILs). 505 epiRILs were obtained by Johannes *et al.* (2009) after four generations of selfing starting from the F_2 generation. Since these 505 lines were (almost) isogenic at the DNA level and differed only in methylation profile, all observable phenotypic variation was then regarded as due to epigenetic and environmental factors, with the impact of genetic polymorphism at the DNA level ruled out. Examining plant height and flowering time, Johannes *et al.* (2009) found that epigenetics contributed approximately 30% of the phenotypic variation. A similar approach was used in Reinders *et al.* (2009) with the only difference being that the genetic polymorphism in the two parental lines was at the *Met1-3* locus, which also has an impact on the whole-genome methylation level. At the end

of the F_8 generation, 68 epiRILs were obtained.

Both Johannes *et al.* (2009) and Reinders *et al.* (2009) found that epigenetic variation contributed a considerable proportion of phenotypic variation, hinting that epigenetic information may help prediction of quantitative traits. When using DNA polymorphisms, whole-genome-enabled prediction models can be viewed as an extension of the single-marker regression models used in GWAS, where instead of finding genomic regions that may be associated with a complex trait, integrating all marker information for prediction and/or artificial selection is the ultimate goal. In a similar context, EWAS studies can also be extended for prediction purposes using data mining and machine learning techniques. Because methylation profiles can explain phenotypic variation and it is widely believed that DNA methylation is the most stable epigenetic modification that can be retained in either mitosis and meiosis, perhaps prediction can be enhanced by using methylation data, as foreseen by González-Recio (2012). In this study, therefore, we used DNA methylation data for building statistical models suitable for prediction purposes, and evaluated how this information could potentially supplement that from DNA polymorphisms.

6.2 Materials and Methods

6.2.1 Data

This study used phenotypic and methylation data. The phenotypic data set is from Johannes *et al.* (2009), and it contains measurements of plant height (PH) and flowering time (FT) collected in two greenhouses for 505 *Arabidopsis* epiRILs and two parental lines. This data was analyzed by Johannes *et al.* (2009) using a mixed effects model, to explore the proportion of phenotypic variance explained by different effects. Their model used greenhouses and micro-environments (i.e., individual planting plots in the greenhouse) as fixed effects and the 505 epiRILs as a random factor. Greenhouse explained 39.61% and 2.45% of phenotypic variance for FT and PH, respectively, and

micro-environment explained 4.12% and 0.086% of phenotypic variance for these two traits; the variance explained by random epiRIL effects accounted for about 30% for both traits. Because the micro-environment arrangement data is no longer available (Johannes, personal communication), we decided to perform the analysis on PH only, as FT was apparently more strongly affected by this factor. The methylation data were downloaded from the Gene Expression Omnibus data repository (access number GSE37284). In this data set, 123 of the 505 epiRILs and the two parental lines were epi-profiled using MeDIP with a customer-designed array chip. Each line was examined at 711,320 probes (loci) located on 5 *Arabidopsis* chromosomes. Each probe is associated with two values: one is the re-scaled $\log_2$ of the signal/background intensity ratio, which describes the enrichment of methylated cytosine proxied by that probe. This information is referred to as methyl-values in subsequent discussion, and a higher methyl-value indicates higher level of methylation. The other value is methylation status (M: methylated; I: intermediately methylated; U: unmethylated) predicted from the methyl-values. Note that the methyl-values are generated from enrichment intensity ratios, so these are relative, rather than absolute, values. Due to this reason, there are typically no threshold values that can be used to perform methylation status calls, and hence the status was predicted using a hidden Markov model (Colomé-Tatché *et al.*, 2012), a commonly used tool in bioinformatic analysis. This predicted methylation status is referred to as methyl-status hereafter. There were no missing values in the methylation data, and after removing epiRILs without phenotypic data, 114 lines remained for subsequent analysis. Therefore, each epiRIL used in this study has 1 phenotypic record on PH and paired methyl-values/methyl-status records at each of 711,320 probes (loci). For more detailed information about the methylation data, see Colomé-Tatché *et al.* (2012) and the NCBI description page. A description of the data processing was given in Cortijo *et al.* (2014a).

6.2.2 Methods and Prediction Models

The methylation data described were used by Cortijo *et al.* (2014b) to map epiQTLs contributing to root length and FT and 3 major epiQTLs were found for both traits. Using analysis of variance, it was found that the broad sense (epi-)heritabilities of these two traits were about 60%, and major epiQTLs explained 87% and 60% of (epi-)heritability in the two traits, respectively. Due to the strong contribution of methylation to variation of phenotype, we decided to explore the predictive power of this information, as suggested by González-Recio (2012). Hence, we built whole (epi-)genome prediction models that are analogous to whole-genome prediction models, where instead of SNP markers, methylation information was used as predictor variables. Most genome-enabled prediction studies (e.g., Meuwissen *et al.*, 2001; de los Campos *et al.*, 2013) use a linear model with the form

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{e}, \tag{6.1}$$

where $\boldsymbol{y}$ is a vector of n phenotypic records; μ is an unknown constant (intercept) common to all individuals; $\boldsymbol{b}$ is a vector of fixed effects with associated incidence matrix $\boldsymbol{X}$; $\boldsymbol{W}$ is an $n \times p$ matrix possessing SNP genotypic codes (e.g., $W_{ij} = 0$, 1, or 2) and $\boldsymbol{\alpha}$ is a $p \times 1$ vector of regression coefficients associated with all SNP loci. Model 6.1 is more statistical than biological when the $\boldsymbol{W}$ matrix contains epigenetic information, compared to when using genomic information such as SNP markers. When performing genomic prediction using SNPs, $\boldsymbol{\alpha}$ represents allelic substitution effects at these marker loci, an important concept in quantitative genetics; hence, a statistical regression coefficient can be linked to a quantitative genetic parameter. Since such concept does not exist in quantitative epigenetic analysis, this implies that a model with this form may not lead itself to interpretability of underlying biological processes. Therefore, we adopted kernel methods for prediction purposes.

Kernel methods: theory

In kernel regression, phenotypes and predictor variables are linked non-linearly, via a kernel function. In a regression problem without nuisance variables, the relationship between an observation y_i and its corresponding covariates $\boldsymbol{x}_i$ is generally written as

$$y_i = g(\boldsymbol{x}_i) + e_i, \tag{6.2}$$

where y_i is the observation on the i^{th} subject and $\boldsymbol{x}_i$ is, say, a $p \times 1$ vector of covariates measured on i; $g(.)$ is some function (usually unknown); e_i is the model residual. For the purpose of describing the kernel methods, it is assumed that phenotypes (y's) and regression covariates ($\boldsymbol{x}$'s) are centered, so Equation 6.2 does not include an intercept. In standard linear regression, $g(\boldsymbol{x}_i)$ is $\boldsymbol{x}_i'\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is a vector of unknown coefficients to be inferred. The most common solution for the weights $\boldsymbol{\omega}$ is obtained by using ordinary least squares (OLS). In whole-genome prediction of complex traits, many Bayesian regression methods use this functional form but assign some penalties to $\boldsymbol{\omega}$ because the "curse of dimensionality" makes OLS not applicable, and often Bayesian techniques are employed (Gianola *et al.*, 2009; Gianola, 2013). The linear additive model often provides a reasonable approximation to the underlying architecture of a complex trait and it is easy to interpret. However, non-additive gene action, for example epistasis, is usually not accounted for, which may lead to incorrect attributions of genetic variation .

One can define $g(\boldsymbol{x}_i) = E(y_i|\boldsymbol{x}_i)$ as the conditional expectation of y_i in Equation 6.2, given $\boldsymbol{x}_i$, which can be inferred using the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964), having the form (Silverman, 1986; Gianola *et al.*, 2006)

$$\hat{g}(\boldsymbol{x}) = \sum_{i=1}^{n} y_i \kappa_h(\boldsymbol{x}_i - \boldsymbol{x}). \tag{6.3}$$

In genome-enabled prediction using high-density markers, n is the number of individuals; $\boldsymbol{x}_i$ is

the $p \times 1$ vector of SNP marker genotypes of individual i; $\boldsymbol{x}$ is the focal point at which the kernel function $\kappa_h(.)$ is evaluated, and h is a smoothing parameter of the kernel function. Because $\boldsymbol{x}_i$ possesses the marker information of individual i, $\kappa_h(\boldsymbol{x}_i, \boldsymbol{x}_j)$ measures the "genomic distance" between individuals i and j by definition. Therefore, the $n \times n$ symmetric matrix $\boldsymbol{\mathcal{K}}_h = \{\kappa_h(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$ measures the pairwise genomic distance of all individuals. According to Gianola and van Kaam (2008), this kernel treatment can be written as (the "dual formulation") the linear regression model

$$\boldsymbol{y} = \boldsymbol{\mathcal{K}}_h \boldsymbol{\alpha} + \boldsymbol{e}, \tag{6.4}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector, $\boldsymbol{\mathcal{K}}_h$ is an $n \times n$ symmetric, positive definite matrix, $\boldsymbol{\alpha}$ is an $n \times 1$ vector of regression coefficients, and $\boldsymbol{e}$ is the model residual with assumption $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$, where σ_e^2 is the residual variance. Under the reproducing kernel Hilbert spaces framework (e.g., Gianola and van Kaam, 2008), one assumes that $\boldsymbol{\alpha}|h \sim N(\boldsymbol{0}, \boldsymbol{\mathcal{K}}_h^{-1}\sigma_{\mathcal{K}}^2)$ and, because $\boldsymbol{\mathcal{K}}_h$ is symmetric and invertible, $\hat{\boldsymbol{\alpha}}$ is estimated as the solution to

$$\left(\boldsymbol{\mathcal{K}}_h + \frac{\sigma_e^2}{\sigma_{\mathcal{K}}^2}\boldsymbol{I}\right)\hat{\boldsymbol{\alpha}} = \boldsymbol{y}. \tag{6.5}$$

Above, $\sigma_{\mathcal{K}}^2$ is the variance captured by the kernel. The vector $\boldsymbol{\mathcal{K}}_h\hat{\boldsymbol{\alpha}}$ estimates the vector of genetic effects marked by SNPs, that is, $g(\boldsymbol{x})$.

Alternatively, starting from $\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{e}$, one can minimize a loss function with form:

$$\ell(\boldsymbol{g}|\lambda) = \|\boldsymbol{y} - \boldsymbol{g}\|^2 + \lambda\|\boldsymbol{g}\|_{\mathcal{H}}^2, \tag{6.6}$$

where λ is a regularization parameter and $\|\boldsymbol{g}\|_{\mathcal{H}}^2$ is the squared norm of $\boldsymbol{g}$ under a Hilbert space $\mathcal{H}$. According to the representer theorem of Kimeldorf and Wahba (1971), the objective function $\boldsymbol{g}$ is reduced to $\boldsymbol{\mathcal{K}}_h\boldsymbol{\alpha}$, as in Equation 6.4, and Equation 6.6 becomes $\ell(\boldsymbol{\alpha}|\lambda) = (\boldsymbol{y} - \boldsymbol{\mathcal{K}}_h\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{\mathcal{K}}_h\boldsymbol{\alpha}) + \lambda\boldsymbol{\alpha}'\boldsymbol{\mathcal{K}}_h\boldsymbol{\alpha}$. When minimizing $\ell(\boldsymbol{\alpha}|\lambda)$ by taking its first derivative with respect to $\boldsymbol{\alpha}$, Equation 6.5 is retrieved if $\lambda = \frac{\sigma_e^2}{\sigma_{\mathcal{K}}^2}$ is assumed. Because optimization of the penalty function is carried out under

a Hilbert space, this approach is known as reproducing kernel Hilbert spaces (RKHS) regression, first proposed in computer sciences and machine learning (Aronszajn, 1950; Kimeldorf and Wahba, 1971; Wahba, 1990, 1999, 2002).

Equation 6.4 has the same form as the "animal" model (e.g., Henderson, 1984; Mrode, 2014) widely used in animal breeding:

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{6.7}$$

where $\boldsymbol{u}$ is the vector of infinitesimal additive effects and $\boldsymbol{Z}$ is the associated incidence matrix. Assumptions for this model are $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{A}\sigma_u^2)$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$, and the best linear unbiased predictor (BLUP) of $\boldsymbol{u}$ can be obtained by solving

$$\left(\boldsymbol{Z}'\boldsymbol{Z} + \frac{\sigma_e^2}{\sigma_u^2}\boldsymbol{A}^{-1}\right)\hat{\boldsymbol{u}} = \boldsymbol{Z}'\boldsymbol{y}, \tag{6.8}$$

where σ_u^2 and σ_e^2 are the additive genetic and residual variances, respectively. Here, the additive relationship matrix $\boldsymbol{A}$ can be interpreted as a kernel matrix measuring the kinship between individuals based on pedigree, as discussed in de los Campos *et al.* (2009) and in Morota and Gianola (2014). Hence, the conventional animal model (pedigree-based BLUP, or P-BLUP) is a special case of RKHS regression. Similarly, the genomic BLUP (G-BLUP) proposed by VanRaden (2008) uses a genomic relationship matrix $\boldsymbol{G} \propto \boldsymbol{X}\boldsymbol{X}'$, with $\boldsymbol{X}$ being the $n \times p$ incidence matrix of marker genotypes, in lieu of the $\boldsymbol{A}$ matrix derived from pedigree. G-BLUP exploits "realized" relationship between individuals using genomic information covering the entire genome. Therefore, G-BLUP is also a special case of RKHS regression. For more details on RKHS regression and its applications to animal breeding, see Gianola *et al.* (2006), Gianola and van Kaam (2008), Gianola and de los Campos (2008), González-Recio *et al.* (2008), de los Campos *et al.* (2009), de los Campos *et al.* (2010), Morota *et al.* (2013), and Morota and Gianola (2014).

In general, the role of a kernel matrix in RKHS regression is to convey similarity between individuals using a certain type of input information, with methylation profiles used here. Although

the choice of the kernel function is arbitrary, as any positive-definite function can be used as a kernel function, multiple factors may affect its choice in practice. For example, the diffusion kernel adopted by Morota *et al.* (2013) has a distance function (Manhattan distance) that may not be optimal for the continuous methylation data. Hence, we chose a Gaussian kernel. By definition, the $(i, j)^{\text{th}}$ element of the Gaussian kernel $\boldsymbol{K}$ is calculated as

$$K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{h}\right), \tag{6.9}$$

where $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the Euclidean distance between vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, in our case the $711,320 \times 1$ vectors of methyl-values of epiRILs i and j, and h is the bandwidth parameter of the kernel, which controls the smoothness of the fitted surface. The choice of the bandwidth parameter is important since it affects the performance of the regression. A number of algorithms have been proposed to optimize the bandwidth parameter (Jones *et al.*, 1996). Here, we determined the optimal bandwidth parameter using a grid search approach under cross-validation, aiming at finding a value that maximized the predictive correlation within a model setting.

From the definition of the Gaussian kernel, all diagonal entries of the kernel matrix are 1, since the Euclidean distance between a vector and itself is always zero. Also, as the distance increases, K_{ij} approaches zero. Hence, the entries of $\boldsymbol{K}$ range between 0 and 1, making the kernel act as a correlation matrix. Therefore, we considered a Pearson's correlation matrix $\boldsymbol{P}$ as a naïve kernel, where $P_{ij} = Corr(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Advantages of using the $\boldsymbol{P}$ matrix are computation-related: 1) it is easy to obtain, and 2) tuning a bandwidth parameter is not needed. Comparisons between prediction performances obtained using the $\boldsymbol{P}$ and the $\boldsymbol{K}$ kernels are described later. A graphical comparison between the $\boldsymbol{P}$ and $\boldsymbol{K}$ kernels is shown in Figure 6.1. In Figure 6.1, the plot at the upper left corner shows the $\boldsymbol{P}$ matrix created from the methyl-values data. Most between-lines correlation range from 0.7 to 0.9 and only few pair-wise correlations are below 0.65. The other three plots represent a $\boldsymbol{K}$ matrix with various h values. It can be seen that h has a big impact on the values of the $\boldsymbol{K}$ matrix. When h is large (1,000,000, upper right corner), the majority of the entries range

from 0.4 to 0.5; for intermediate h (500,000, lower left corner), most entries are between 0.2 to 0.7; when h is small (250,000, lower right corner), almost all entries are smaller than 0.5 except for the diagonal elements.

Given a kernel $\boldsymbol{\mathcal{K}}$ and a vector of fixed effects $\boldsymbol{b}$ (in our case the greenhouses only), the prediction model can be written in matrix form as:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\mathcal{K}}\boldsymbol{\alpha} + \boldsymbol{e}, \tag{6.10}$$

where $\boldsymbol{y}$ is the vector of phenotypic records (PH here); μ is an unknown intercept common to all observations; $\boldsymbol{X}$ is the incidence matrix of fixed greenhouse effects; $\boldsymbol{\alpha}$ is the random vector of regressions on the kernel associated with epigenetic variation, with assumed distribution $N(\mathbf{0}, \boldsymbol{\mathcal{K}}^{-1}\sigma^2_{\mathcal{K}})$, where $\sigma^2_{\mathcal{K}}$ is a variance component associated with the kernel; and $\boldsymbol{e}$ is the model residual with distribution $N(\mathbf{0}, \boldsymbol{I}\sigma^2_e)$.

Prediction using pre-selected probes

Our main goal is to build prediction models using epigenetic information as a potential supplement to genomic variation (e.g., SNP markers), as foreseen by González-Recio (2012). In animal and plant breeding, a training population with thousands of individuals is usually needed. However, methylation profiling experiments are extremely expensive, at least at present. Thus, cost is usually a main consideration in epigenetic studies, and data sets with hundreds of profiled individuals are commonly viewed as large scale experiments. Thanks to Meissner *et al.* (2005), a molecular genetic technique called reduced representation bisulfite sequencing (RRBS) has been used to take only a small subset of all available probes as proxies to describe the methylation level of the whole genome, which may reduce experimental costs drastically and make experiments executed on a larger cohort possible. According to the mechanisms of DNA methylation known so far, cytosine in a CpG dinucleotide context (cytosine followed by guanine, "p" indicates the phosphate bond

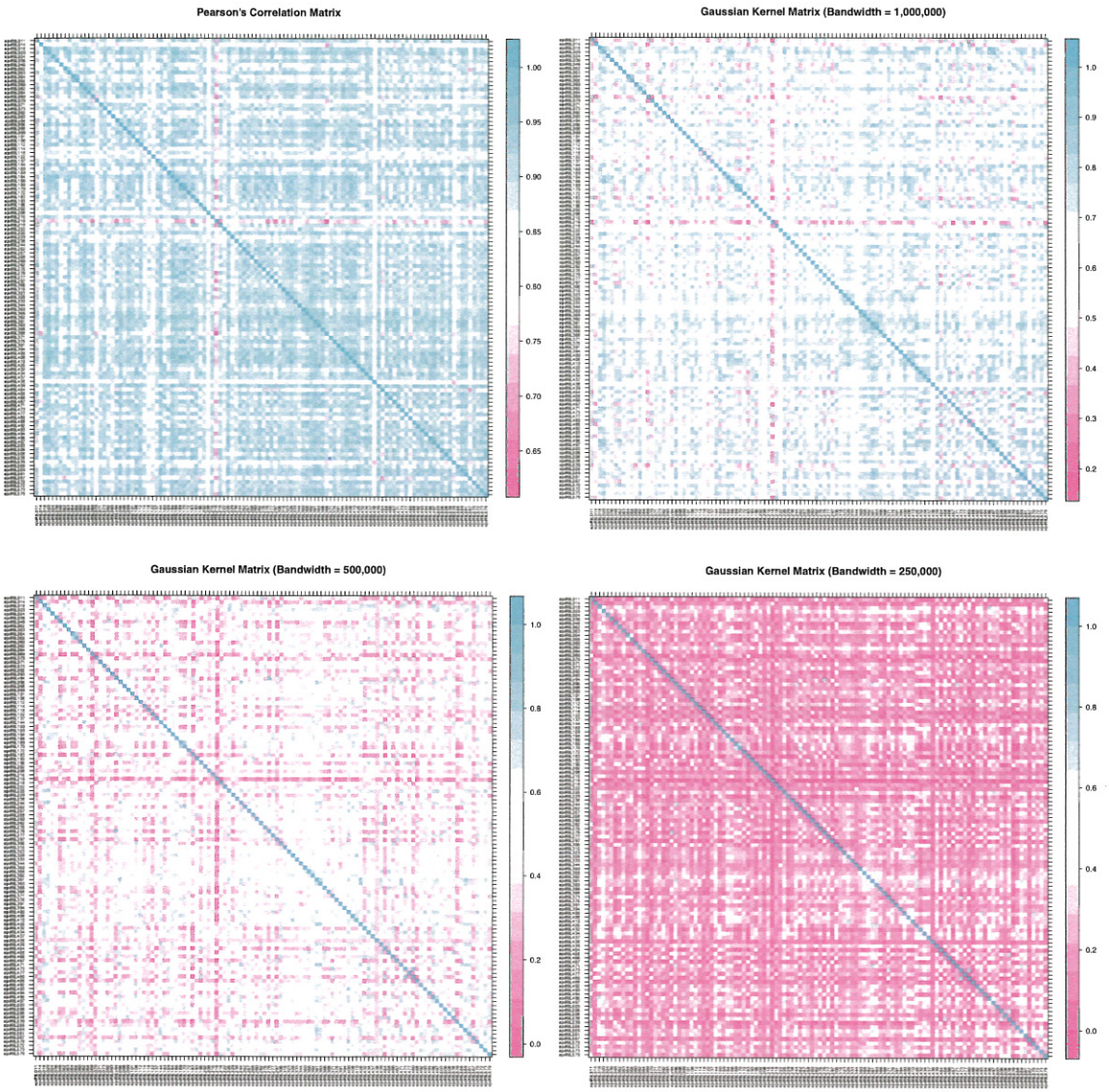


Figure 6.1: Visualization of several kernel matrices. The four matrices displayed are Person's correlation matrix (top left); and Gaussian kernels with bandwidth parameters equal to 1,000,000 (top right), 500,000 (bottom left), and 250,000 (bottom right).

in between) is the main target of DNA methylation in eukaryotic cells. Thus, genomic regions with high CpG content may represent the methylation profile of the entire genome and hence are chosen for BS-Seq in RRBS (note that CpG content is different from CG content; the latter evaluates cytosine and guanine frequencies separately). In the mouse, according to Meissner *et al.* (2005), these selected regions comprised only $\sim$12 Mb of the whole genome (less than 0.5%), but

captured most variation at the methylation level. This suggests that a subset of representative probes may perhaps provide a similar predictive performance to that from all probes. If this is the case, prediction using representative probes would be less expensive and computing burden would be lessened because generating a kernel matrix is potentially time-consuming.

Considering the size of the murine ($\sim 3,000$ Mb) and of the *Arabidopsis* (~ 120 Mb) genomes, we decided to select the top (see below) 10% of the profiled probes in *Arabidopsis* such that the genomic regions in which these probes reside had about 12 Mb in total. Thus, the cost needed for the experiment would not exceed the magnitude of what was suggested by RRBS in mouse. The criterion for this selection was based on the observed/expected (O/E) CpG ratio defined (Gardiner-Garden and Frommer, 1987) as

$$\frac{\text{Number of CpG}}{\text{Number of C} \times \text{Number of G}} \times \text{Total number of nucleotides in the sequence,}$$

which is a statistic describing the frequency of occurrence of CpG dinucleotides. Besides CpG dinucleotides, it has been found that trinucleotides CpHpG and CpHpH (H = A, C, or T) are target sites of DNA methylation in plants as well (Henderson and Jacobsen, 2007; Lister *et al.*, 2008). Thus, we also calculated the O/E ratio for these two trinucleotides. According to the reference genome (TAIR7, downloaded from `http://www.arabidopsis.org`), the total length of the *Arabidopsis* genome is 119,186,497 bp. With 711,320 probes on the designed chip, on average there is 1 probe for every 167 base pairs. The average length of all probe sequences is 55.2 bp (max 75 bp, min 50 bp), which means that the DNA segment between two probes is roughly 112 bp long, on average. Considering that 55 bp may not be an adequate length for calculating the O/E ratio with accuracy, especially for the two trinucleotides, we decided to extend the region of examination by 120 bp to the upstream of each probe. After this extension, the estimation of O/E CpG ratio is expected to be more accurate, and the number 120 was chosen because: 1) it fills the gap between two probes, so this assures that the whole genome is under examination, and 2) the overlap between adjacent probes after extension is reduced.

In order to make up 10% of total probes, we chose the top 5% probes with highest O/E ratio for CpG dinucleotides, top 2.5% probes with highest O/E ratio for either CpHpG or CpHpH. This 2:1:1 partition comes from the fact that in *Arabidopsis*, the fractions of ^{m5}C identified in CpG, CpHpG, and CpHpH contexts are about 55%, 23% and 22%, respectively (Lister *et al.*, 2008). As a result, we selected 35,585, 17,783, and 17,783 probes based on the CpG, CpHpG, and CpHpH contents, respectively, and ended up with 65,506 probes (9.2% of all probes) in total (with some overlap between different contents). After mapping back to the genome annotation file (TAIR7, downloaded from `http://www.arabidopsis.org`), it was found that within these 65,506 probes, 10,044 (15.3%) were located in promoter regions of genes, and these 10,044 probes cover 40.9% of total promoters; 12,074 (18.4%) were found in protein-coding regions (CDS); 2,329 and 2,005 (3.6% and 3.1%) were in 5' UTR and 3' UTR regions, respectively; and 2,534 (3.9%) probes were in pseudogenic exons. Also, 1,418 and 16,258 (2.2% and 24.8% of the 65,506 pre-selected probes) were found in the intron and transposon regions, respectively, with the reference information provided by Cortijo *et al.* (2014a). Lastly, 18,620 (28.4%) probes did not map to any annotated region according to the current annotation file. A graphical representation of the distribution of selected probes by genomic element groups is shown in Figure 6.2. In addition to a model using all available probes, a prediction model using these 65,506 pre-selected probes was built as well.

In their RRBS study, Meissner *et al.* (2005) reported that the representative subset covered $> 90\%$ gene promoter regions, while in our bioinformatic search, only 40% of the promoter regions were covered by pre-selected probes as described above. This is probably due to differences between species since the original RRBS method was developed in the mouse. Given the important role of gene promoter regions in epigenetic regulation of gene expression, we attempted selecting a subset of probes with a different criterion such that more promoter regions could be covered. In epigenetics, CpG islands (CGIs) are CpG-rich regions that are usually unmethylated and located in the gene promoter region. In humans, at least $60 \sim 70\%$ gene promoter regions overlap with CGIs (Illingworth and Bird, 2009). CGI shores are close proximity regions ($\sim$ 2 kb of both upstream or

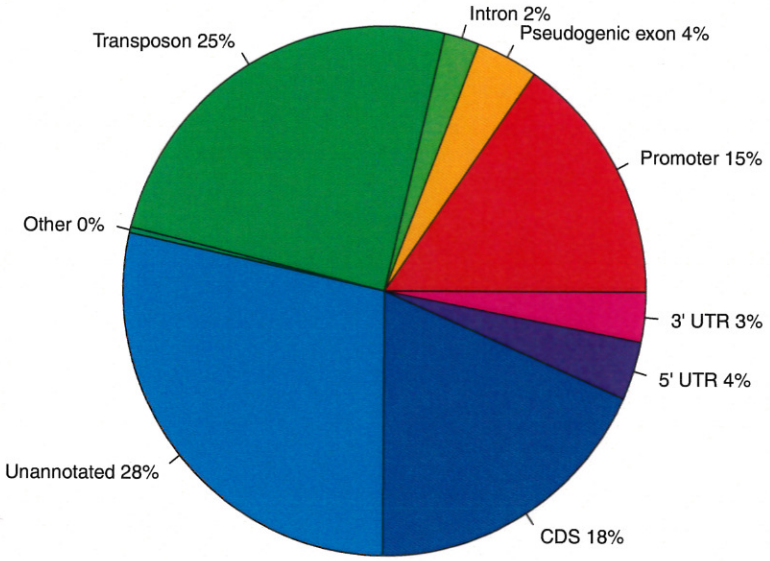


Figure 6.2: Distribution of the selected 65,506 probes. Most selected probes (about 70%) were located in annotated regions (e.g., CDS, 5' UTR, 3' UTR, promoter, intron, etc.).

downstream) of CGIs (Portela and Esteller, 2010). Recent studies suggested that 70% of differentially methylated regions in epigenetic reprogramming are associated with CGI shores (Doi *et al.*, 2009; Ji *et al.*, 2010). Therefore, probes located in CGI shores were also selected, and these probes may constitute another (independent) subset that represents the whole-genome methylation profile. Following the definition of CGI given by Gardiner-Garden and Frommer (1987), we found 23,640 CGIs, and the probes located in the shore regions of these CGIs covered 65.6% of all promoters (Table 6.1). Thus, apart from different kernel matrices applied, prediction was performed using 1) all probes available in the data set; or 2) pre-selected probes based on CpG/CpHpG/CpHpH contents (referred to as Contents Rule hereafter); or 3) pre-selected probes located in the CGI shore region (referred to as CGI Rule hereafter).

Prediction using methyl-status data

When P-BLUP and G-BLUP are viewed as kernel methods, the $\boldsymbol{A}$ and $\boldsymbol{G}$ kernel matrices have an explicit biological meaning. For example, the kinship matrix $\boldsymbol{A}$ reflects the expected fraction

Table 6.1: Number of promoters covered by CGI (definition of Gardiner-Garden and Frommer, 1987) shores.

	No. promoter	No. CGI	No. promoter covered by CGI shore
Chr1	6354	1716	4111 (64.7%)
Chr2	3990	1063	2583 (64.7%)
Chr3	4902	1363	3289 (67.1%)
Chr4	3621	1025	2402 (66.3%)
Chr5	5677	1624	3709 (65.3%)
Total	24544	6791	16094 (65.6%)

of IBD alleles shared by a pair of relatives and the $\boldsymbol{G}$ matrix can be viewed as a realization of $\boldsymbol{A}$ given the observed molecular markers, or as a "molecular similarity matrix" based on the DNA polymorphisms. Thus, variance components associated with $\boldsymbol{A}$ or $\boldsymbol{G}$ have a clear genetic basis. The correlation matrix $\boldsymbol{P}$ and any of the Gaussian kernels with specific bandwidth parameters used here, on the other hand, are constructed from methylation profiles and reflect only epigenetic similarity in some manner. Hence, variance components associated with these kernels do not have an easy biological interpretation except that of measuring a contribution to phenotypic variance. Further, when a Gaussian kernel $\boldsymbol{K}$ is used, the bandwidth parameter h has a large impact on the values in $\boldsymbol{K}$, as depicted in Figure 6.1. As such, one may expect that various distinct $\hat{\sigma}^2_{\mathcal{K}}$ will be obtained when different values are set to h; hence, $\hat{\sigma}^2_{\mathcal{K}}/(\hat{\sigma}^2_{\mathcal{K}}+\hat{\sigma}^2_e)$ will vary as well. In order to obtain a more meaningful partition of phenotypic variation explained by epigenetic polymorphisms, we built an additional kernel matrix for RKHS.

Because the methylation state of one copy at a single locus (e.g., an epi-allele or a single cytosine at a CpG dinucleotide) can be only methylated or unmethylated, the absolute methylation level β (as obtained by BS-Seq, for example) is always measured as a ratio that ranges between 0 and 1, with the numerator being the number of methylation incidences in a sample. Under some circumstances, methylation at the locus under investigation can be classified into one of the three categories: methylated (M), intermediately methylated (I), or unmethylated (U), according to the β value at that locus (Meissner *et al.*, 2008; Du *et al.*, 2010). If two DNA segments with similar nucleotide sequence but different methylation status (e.g., one is methylated and the other is not)

are considered as two epi-alleles, this classification provides an approximation to the underlying "epi-genotypes" such that M and U stand for the "epi-homozygotes" for one of the two epi-alleles and I is the "epi-heterozygote". Analogous to the SNP coding system, we can use 2, 1, and 0 to code M, I, and U, and generate a kernel matrix mimicking the $\boldsymbol{G}$ matrix in G-BLUP (VanRaden, 2008), which we call the epi-$\boldsymbol{G}$ matrix. This required little extra effort since methyl-status was available in the methylation data set. However, this approach has some pitfalls: 1) when continuous methyl-values are converted to discrete methyl-status, information is lost; 2) once a numeric coding is arrived at, many probes would be excluded from downstream analysis because their "epi-MAF" would be lower than 0.05 (MAF stands for "minor allele frequency"). In the current data set, only 206,600 probes were kept for subsequent analysis after this epi-MAF filtering. Nevertheless, this epi-$\boldsymbol{G}$ kernel may be more biologically intuitive than a Gaussian kernel generated from methyl-values since the numeric coding used to generate the epi-$\boldsymbol{G}$ kernel is an absolute count of a certain epi-allele of an epi-genotype. Thus, a prediction model can be built and implemented as in G-BLUP, and the variance component associated with epi-$\boldsymbol{G}$ would estimate the proportion of total variance explained by epigenetic variation with a clearer biological sense.

6.3 Results

6.3.1 Predictions with Different Kernels

Considering that the data set has only 114 epiRILs, we used a leave-one-out cross-validation (LOO CV) for model evaluation throughout the study. When the correlation matrix $\boldsymbol{P}$ was used as a naïve kernel, the predictive correlation was 0.384. When a Gaussian kernel was used, the predictive correlation varied according to the bandwidth parameter chosen. In this case, when all probes were used to create the kernel matrix, the best prediction performance was obtained when the bandwidth parameter was set to 140,000, and the predictive correlation was 0.531, with predictive mean squared error (MSE) equal to 32.16.

It can be seen that a reasonable predictive correlation was reached when using the Gaussian kernel, which performed much better than the correlation kernel. However, the bandwidth parameter played an important role on model performance (Figure 6.3). Taking the four kernels in Figure 6.1 as an example, $\boldsymbol{P}$ had entries ranging from about 0.6 to 1, which means that the "dissimilarity" between each line must be distinguished within a 0.4 range. On the other hand, all three Gaussian kernels in Figure 6.1 ranged from about 0 to 1, such that pair-wise dissimilarity was better distinguished on a wider scale. Thus, it was not surprising that the Gaussian kernel outperformed the correlation kernel. Predicting an unobserved record borrows information from observations on similar individuals. Thus, the "similarity" between lines matters. From the definition of the kernel matrix (Equation 6.9), the off-diagonal elements are proportional to the exponential of the bandwidth parameter, and all are close to zero if h is small (Figure 6.4). This makes the kernel matrix to be "confounded" with the identity matrix, which represents the variance-covariance structure of the model residuals. According to de los Campos *et al.* (2010), this type of kernel matrix captures "local" similarity, focusing mainly on the comparison of an individual with itself and few other individuals with highest similarities. A "global" kernel with a larger bandwidth parameter, on the other hand, will also take into account comparisons between more (epi-)genetically distant individuals. Therefore, the "optimal" bandwidth parameter should provide a balance between local and global comparisons between different lines, using the available data. Unless multiple kernels with different bandwidth parameters are fitted simultaneously (e.g., Tusell *et al.*, 2014), a kernel with an intermediate h is expected to provide the best predictive correlation (Figure 6.3, black solid line). A similar pattern was observed for predictive MSE (Figure 6.3, blue dotted line).

6.3.2 Prediction Using Pre-selected Probes

For pre-selected probes, models using different kernel matrices were evaluated as well; again, the bandwidth parameter for the Gaussian kernel was determined based on a grid search via LOO CV. When using the $\boldsymbol{P}$ kernel, the predictive correlation for Contents rule and CGI rule probes were

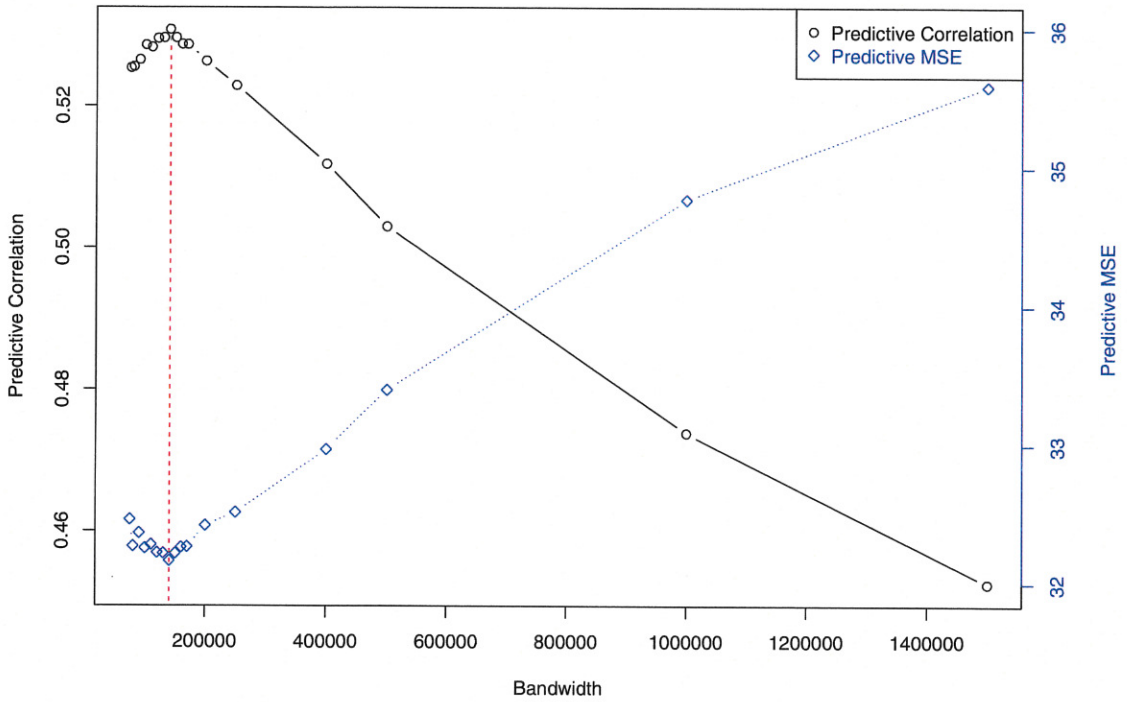


Figure 6.3: Predictive correlation and predictive MSE with various bandwidth parameters under LOO CV (all probes used). The kernel with the highest predictive correlation and lowest MSE is denoted by dashed red line, where bandwidth is 140,000.

0.398 and 0.395, respectively, slightly higher than when all probes were used for prediction. When a Gaussian kernel was used, the highest predictive correlations for these two sets of pre-selected probes were 0.532 and 0.531, respectively, given an appropriate bandwidth parameter. This result was the same as when all probes were used (Table 6.2, Figure 6.5). As a comparison, we also drew 10 subsets of probes, each consisting of a random 10% of all available probes, to evaluate the usefulness of pre-selection of representative probes according to different criteria. Results showed that the predictive correlations using randomly selected probes were all lower than when using representative probes selected according to an explicit criteria, regardless of the kernel used in prediction.

Our results suggest that a properly selected subset of all probes is able to capture most variation at the methylome level. Therefore, prediction of a larger cohort with a limited budget is possible since only a small fraction of "loci" is needed for methylation profiling with computation time

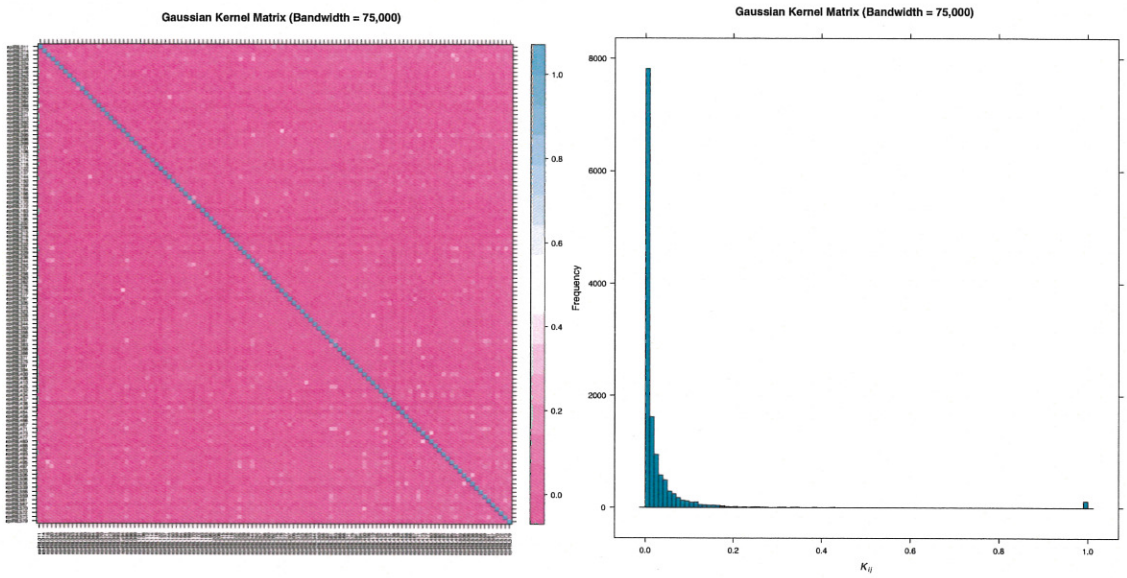


Figure 6.4: A Gaussian kernel with small bandwidth parameter is similar to an identity matrix. The left panel is a visualization of a Gaussian kernel with bandwidth equal to 75,000, which makes it closer to an identity matrix than any of the other matrices in Figure 6.1. The right panel is a histogram of the values in this kernel matrix showing that most values in this matrix are either exactly zero or very close to zero. A short bar at $K_{ij} = 1$ represents the diagonal elements of this kernel matrix.

Table 6.2: Comparison between prediction results using all probes and pre-selected probes.

Kernel		All Probes	Contents Rule Probes	CGI Rule Probes
Correlation	$Corr(\boldsymbol{y}, \hat{\boldsymbol{y}})$	0.384	0.398	0.395
	MSE	38.28	37.73	37.83
Gaussian	$Corr(\boldsymbol{y}, \hat{\boldsymbol{y}})$	0.531	0.532	0.531
	MSE	32.16	32.08	32.13

decreasing drastically. This could be very useful in livestock or crop production since there are usually thousands of individuals in a breeding program that need to be chipped (i.e., methylation profiled), which is still very costly. Lower predictive correlations obtained using randomly selected probes indicated that the most relevant methylation variation is harbored in previously identified regions, i.e., high CpG/CpGpH/CpHpH content regions or CGI shore regions.

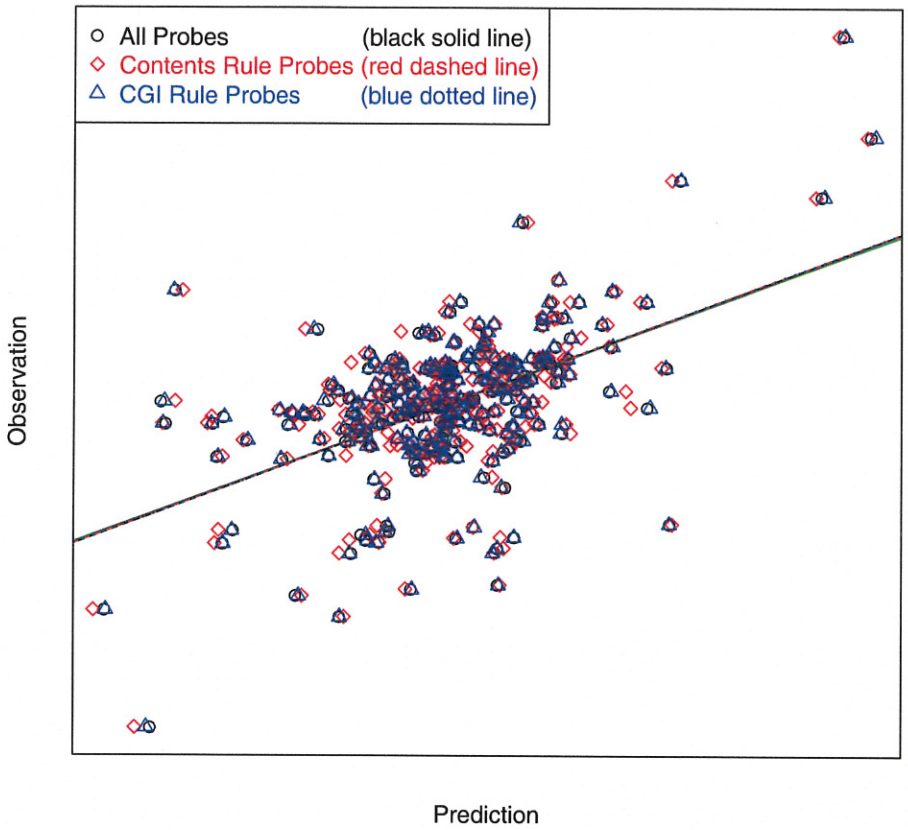


Figure 6.5: Graphical representation of prediction performance using different set of probes in a Gaussian kernel. The green solid line is a 45° line passing through the origin, the other three lines are fitted lines of regressing observation on predictions. No differences were observed for the three different sets of probes used in prediction.

6.3.3 Prediction Using the epi-G Kernel

When using epi-$\boldsymbol{G}$ in prediction, the predictive correlation was 0.505 when all probes were used; predictive correlations were 0.494 and 0.507, if probes were pre-selected based on the Contents or the CGI rules, respectively. Thus, the predictive correlation using epi-$\boldsymbol{G}$ was somewhat lower than that when a Gaussian kernel was used, perhaps due to the loss of information from discretization of methyl-values. The estimated variance component associated with epi-$\boldsymbol{G}$ was 41.57 (SD = 1.69 under cross-validation) and the residual variance was estimated as 22.75 (SD = 0.45), so the proportion of total variance explained by epi-genotype was 0.646, close to what was reported in Cortijo *et al.* (2014b). This proportion was 0.656 and 0.647 when only pre-selected probes (with two

criteria, respectively) were used (Table 6.3). When using Gaussian kernels, on the other hand, the variance component associated with the kernel matrix represented 0.542 of the phenotypic variance (all probes used, bandwidth = 140,000), which was lower than with the epi-$\boldsymbol{G}$ kernel.

Arguably, it is difficult to assess which kernel provides a more meaningful proportion of phenotypic variance explained by the methylation profile, since the true variance components are unknown. The bandwidth parameter had a strong impact on predictive correlation and on the estimated variance components as well (Figure 6.6). The estimated variance components associated with the Gaussian kernel were very large when the bandwidth parameter was large, and the proportion of phenotypic variation explained by the kernel matrix seemed excessive (up to 0.85). Note that the residual variance was essentially independent of the bandwidth parameter value. Obviously, there are problems here. When using the epi-$\boldsymbol{G}$ kernel, on the other hand, the proportion of phenotypic variation explained by epigenetic polymorphisms seemed more reasonable, and predictions obtained using this kernel gave a better predictive correlation than when using the $\boldsymbol{P}$ kernel ($Corr(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 0.384$). Also, the regression of testing set observations on predicted values was 0.99, much higher than for $\boldsymbol{P}$ kernel ($b_{\boldsymbol{y},\hat{\boldsymbol{y}}} = 0.90$, Figure 6.7).

Table 6.3: Estimated variance components associated with a Gaussian and an epi-$\boldsymbol{G}$ kernel.

Kernel		All Probes	Contents Rule Probes	CGI Rule Probes
Gaussian	$Corr(\boldsymbol{y}, \hat{\boldsymbol{y}})$	0.531	0.532	0.531
	$\hat{\sigma}^2_{\mathcal{K}}$	25.59	23.75	25.61
	$\hat{\sigma}^2_e$	21.59	21.44	21.58
	$\hat{\sigma}^2_{\mathcal{K}}/(\hat{\sigma}^2_{\mathcal{K}} + \hat{\sigma}^2_e)$	0.542	0.525	0.543
epi-$\boldsymbol{G}$	$Corr(\boldsymbol{y}, \hat{\boldsymbol{y}})$	0.505	0.494	0.507
	$\hat{\sigma}^2_{\mathcal{K}}$	41.57	44.45	41.58
	$\hat{\sigma}^2_e$	22.75	23.35	22.70
	$\hat{\sigma}^2_{\mathcal{K}}/(\hat{\sigma}^2_{\mathcal{K}} + \hat{\sigma}^2_e)$	0.646	0.656	0.647

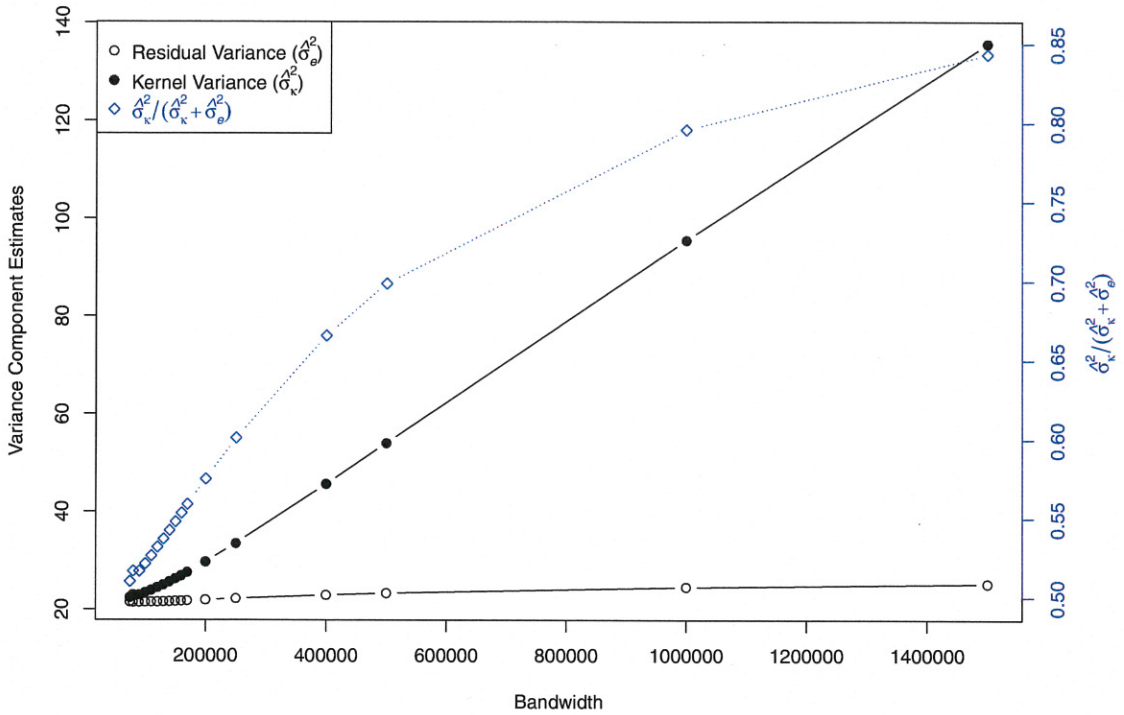


Figure 6.6: Estimated variance components (black lines, left y-axis) and variance component ratio (blue dotted line, right y-axis) with various bandwidth parameters of the Gaussian kernel (all probes used).

6.4 Discussion

6.4.1 Prediction Using Epigenomic Data

It was found that using methylation data for prediction of plant height in *Arabidopsis* produced a reasonable predictive correlation. The kernel matrix used reflects epigenetic similarity between epiRILs based on their methylation profiles. Thus, epigenomic information may complement genomic information at the DNA level. The predictive correlation and mean squared error values were maintained when only pre-selected probes were used. Hence, use of representative probes may be a good alternative to reduce the cost of methylation profiling and computing time as well.

Non-parametric prediction using kernel methods is relatively simpler than Bayesian regression models based on Markov Chain Monte Carlo with an enormous number of proposal distributions.

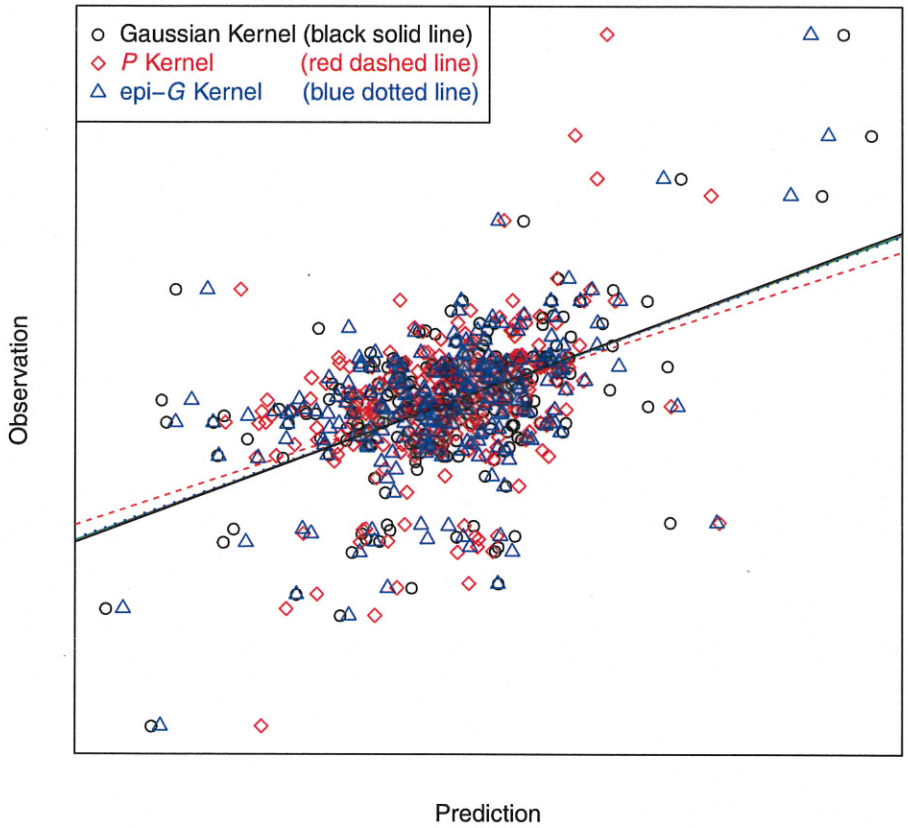


Figure 6.7: Graphical representation of prediction performance using different kernel matrices (all probes used). The green solid line is a 45° line passing through the origin. Regression lines using the Gaussian kernel and the epi-$\boldsymbol{G}$ kernel overlapped and were close to the 45° line, whereas the line using $\boldsymbol{P}$ kernel had a smaller slope.

However, in most cases the variance components associated with a kernel matrix do not provide a meaningful explanation of the underlying biological processes, except for P-BLUP and G-BLUP that are two special cases of RKHS regression. It was observed that when a Gaussian kernel was used for prediction, estimated variance component associated with the kernel varied much with different choices of the bandwidth parameter h, probably due to the big impact of h on the values of the kernel matrix. This variation produced a wide range of $\hat{\sigma}^2_{\mathcal{K}}/(\hat{\sigma}^2_{\mathcal{K}} + \hat{\sigma}^2_e)$ ratios, making it difficult to assess phenotypic variance explained by epigenetic polymorphisms (Figure 6.6), even though the best predictive performance was obtained when a Gaussian kernel was used. In order to cope with this difficulty, an epi-$\boldsymbol{G}$ that mimicking the $\boldsymbol{G}$ matrix in G-BLUP was used as a kernel in RHKS regression. Since the epi-$\boldsymbol{G}$ kernel was generated from discrete methyl-status that

was converted from continuous methyl-values data, a reduced predictive performance was observed probably due to loss of information during this data conversion process. However, the methyl-status data approximates the underlying epi-genotypes of each loci, and hence the variance component associated with the epi-$\boldsymbol{G}$ is interpreted as in G-BLUP, which is clearly based on a biological concept.

Despite the potential usefulness of epigenetic information in phenotype prediction suggested by our results, it should be noted that DNA methylation is reversible (i.e., a methylated DNA molecule can be de-methylated). Hence, methylation data is relatively unstable compared to DNA polymorphisms. The reversibility of DNA methylation may produce "epimutation" events (Becker *et al.*, 2011). Therefore, the entire methylome represents the dynamics of epimutations, and a particular methylation dataset should be viewed as a "snapshot" of the methylome at a specific time from a specific tissue. In order to enhance phenotypic prediction performance further, information from multiple snapshots could be useful. Although methylation profiling is still expensive, its cost has decreased in recent years, and this trend is expected to continue.

6.4.2 Integrating Genomic and Epigenomic Data in Prediction

Previous results suggested that epigenetic information can be used alone for whole-genome prediction of plant height, as a reasonable prediction performance was obtained. Therefore, it should be possible to use epigenetic information in addition to DNA information for the same purposes. For example, Vázquez *et al.* (2014) showed that the inclusion of multi-layer -omics data in human epidemiology can increase the predictive correlation of disease risk drastically. The most obvious explanation is that some layers are "closer" to the ultimate phenotype than the DNA layer. Further, fitting several kernels together tends to enhance predictive performance. For example, fitting a pedigree-based relationship matrix (the $\boldsymbol{A}$ matrix) and a genome-based relationship matrix (the $\boldsymbol{G}$ matrix) together can give a higher predictive correlation than when only one matrix was fitted (Crossa *et al.*, 2010). $\boldsymbol{A}$ and $\boldsymbol{G}$ may provide information from different perspectives, with

$\boldsymbol{G}$ supplementing information that is not captured by $\boldsymbol{A}$. Also, a prediction model with both $\boldsymbol{A}$ and $\boldsymbol{G}$ is a model with multiple kernels when viewing P-BLUP and G-BLUP as special cases of RKHS regression. Tusell *et al.* (2014) reported that a multi-kernel model performed better than a single-kernel model, as anticipated by de los Campos *et al.* (2010). The benefit from fitting multiple kernels simultaneously can be enhanced if all kernels are mutually orthogonal (Morota *et al.*, 2014). A recent study showed that genetic and epigenetic information can be uncoupled by epimutation over an evolutionary time scale (van der Graaf *et al.*, 2015), so a higher predictive correlation could be expected when information from the epigenome is included in a prediction model, as suggested by Vázquez *et al.* (2014), since this extra information is distinct from the information conveyed by DNA polymorphisms.

Biologically, this phenomenon can be interpreted as follows. The DNA sequence is transcribed into RNA and subsequently translated into protein, the building blocks of final phenotypes. Therefore, information at the protein layer (proteome) is "closest" to and genomic information is most "distant" from phenotypes in this biological pathway. Hence, proteomic information might provide better predictions of the phenotypes than genomic information. Similarly, the epigenomic information, which lies between that conveyed by DNA and RNA layers, should be more informative, if available. However, exploration at the DNA level is crucial for artificial selection and the epigenetic data would be informative only if transmission between generation is verified. This does not contradict the argument that using information from multiple layers is expected to give stronger predictive correlations, as indicated by González-Recio (2012) and corroborated by Vázquez *et al.* (2014).

In short, including both DNA and epigenetic information into a prediction model may be fruitful. For example, when the epi-$\boldsymbol{G}$ kernel is used along with a $\boldsymbol{G}$ matrix (using SNP data), the estimated variance components should help interpretation of the proportion of phenotypic variance attributed to genetic and epigenetic variation. Also, perhaps the loss of information incurred when forming the epi-$\boldsymbol{G}$ kernel might be compensated by $\boldsymbol{G}$. Therefore, using epigenomic and genomic

information together may have potential and additional study is needed.

6.5 Conclusion

We built prediction models non-parametrically using DNA methylation data. We chose RKHS regression for prediction because, unlike with prediction using SNP data, estimated regressions using methylation data do not have an obvious interpretation that links to model parameters via some biological concept. In RKHS regression, a kernel matrix described epigenetic similarities between different epiRILs. The tuning procedure is easier than for a parametric model, where a Bayesian treatment and MCMC techniques are usually needed.

We used different kernels in RKHS regression, namely the naïve correlation matrix $\boldsymbol{P}$ and a Gaussian kernel $\boldsymbol{K}$ with different bandwidth parameters. When the bandwidth parameter was selected appropriately, the model with a Gaussian kernel performed better than that with a $\boldsymbol{P}$ kernel. Since a reasonably good predictive correlation was observed, this suggested that epigenetic information may be useful in whole-genome prediction as a source of information that does not reside in a DNA sequence. Furthermore, the value of the predictive correlation was retained when using pre-selected representative probes, suggesting an avenue for cost reduction in prediction studies.

The performance of RKHS regression with a Gaussian kernel was strongly affected by its associated bandwidth parameter, not only in terms of the predictive correlation and predictive mean squared error, but also with respect to the variance component associated with the kernel matrix. This is because epigenetic similarities between individuals provided by the kernel matrix are based on a relative metric, instead of an absolute one. Therefore, the proportion of variance explained by the kernel does not give a meaningful interpretation of the proportion of phenotypic variance explained by epigenetic variation. On the other hand, a kernel matrix created from coded methylation status (epi-$\boldsymbol{G}$) mimicked the genomic relationship matrix $\boldsymbol{G}$, and gave an estimated proportion

of total variance explained by epigenetic variation of around 65%. Although a small degradation in prediction performance is incurred when this epi-$\boldsymbol{G}$ kernel is applied, a better understanding of the importance of epigenetic variance can be obtained.

Using epigenetic information in addition to DNA polymorphisms in prediction has been studied by other authors in human epidemiology (e.g., Vázquez *et al.*, 2014), and results have suggested that this extra information may lead to a pronounced impact on prediction performance. Based on their results, and on the empirical observation that RKHS regression with multiple kernels performs better than a single-kernel regression (Tusell *et al.*, 2014), we conclude that inclusion of epigenetic information in prediction models may be warranted, and possibly useful in livestock and crop production, as suggested by González-Recio (2012).

References

Aronszajn, N., 1950. Theory of reproducing kernels. T. Am. Math. Soc., 68: 337–404

Becker, C., J. Hagmann, J. Müller, *et al.*, 2011. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature, 480(7376): 245–249

Bell, C. G., 2013. Epigenome-Wide Association Studies: Potential Insights into Human Disease. In A. Naumova and C. Greenwood (Editors), *Epigenetic and Complex Traits*, pp. 287–317. Springer New York

Berger, S. L., 2002. Histone modifications in transcriptional regulation. Curr. Opin. Genet. Dev., 12(2): 142–148

Bird, A. P., 1984. DNA methylation versus gene expression. J. Embryol. Exp. Morphol., 83 (Suppl.): 31–40

Cassidy, S. B., E. Dykens, and C. A. Williams, 2000. Prader-Willi and Angelman syndromes: sister imprinted disorders. Am. J. Med. Genet., 97(2): 136–146

Cheung, P. and P. Lau, 2005. Epigenetic regulation by histone methylation and histone variants. Mol. Endocrinol., 19(3): 563–573

Colomé-Tatché, M., S. Cortijo, R. Wardenaar, *et al.*, 2012. Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proc. Natl. Acad. Sci. U.S.A., 109(40): 16240–16245

Cortijo, S., R. Wardenaar, M. Colome-Tatche, *et al.*, 2014a. Genome-wide analysis of DNA methy-

lation in *Arabidopsis* using MeDIP-chip. In C. Spillane and P. C. McKeown (Editors), *Plant Epigenetics and Epigenomics: Methods and Protocols*, Methods in Molecular Biology (vol. 1112), pp. 125–149. Humana Press

Cortijo, S., R. Wardenaar, M. Colome-Tatche, *et al.*, 2014b. Mapping the epigenetic basis of complex traits. Science, 343(6175): 1145–1148

Crossa, J., G. de los Campos, P. Pérez, *et al.*, 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics, 186(2): 713–724

de los Campos, G., D. Gianola, and G. J. Rosa, 2009. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J. Anim. Sci., 87(6): 1883–1887

de los Campos, G., D. Gianola, G. J. Rosa, *et al.*, 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. (Camb), 92(4): 295–308

de los Campos, G., J. M. Hickey, R. Pong-Wong, *et al.*, 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193(2): 327–345

Doi, A., I. H. Park, B. Wen, *et al.*, 2009. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat. Genet., 41(12): 1350–1353

Du, P., X. Zhang, C. C. Huang, *et al.*, 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics, 11: 587

Egger, G., G. Liang, A. Aparicio, *et al.*, 2004. Epigenetics in human disease and prospects for epigenetic therapy. Nature, 429(6990): 457–463

Esteller, M., 2008. Epigenetics in Cancer. N. Engl. J. Med., 358(11): 1148–1159

Flanagan, J. M., 2015. Epigenome-wide association studies (EWAS): past, present, and future. In M. Verma (Editor), *Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment, and Prognosis*, Methods in Molecular Biology (vol. 1238), pp. 51–63. Humana Press

Frommer, M., L. E. McDonald, D. S. Millar, *et al.*, 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. U.S.A., 89(5): 1827–1831

Gardiner-Garden, M. and M. Frommer, 1987. CpG islands in vertebrate genomes. J. Mol. Biol., 196(2): 261–282

Gianola, D., 2013. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics, 194(3): 573–596

Gianola, D. and G. de los Campos, 2008. Inferring genetic values for quantitative traits non-parametrically. Genet. Res. (Camb), 90(6): 525–540

Gianola, D., G. de los Campos, W. G. Hill, *et al.*, 2009. Additive genetic variability and the Bayesian alphabet. Genetics, 183(1): 347–363

Gianola, D., R. L. Fernando, and A. Stella, 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics, 173(3): 1761–1776

Gianola, D. and J. B. C. H. M. van Kaam, 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics, 178(4): 2289–2303

González-Recio, O., 2012. Epigenetics: a new challenge in the post-genomic era of livestock. Front. Genet., 2: 106

González-Recio, O., D. Gianola, N. Long, *et al.*, 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics, 178(4): 2305–2313

Henderson, C. R., 1984. *Applications of Linear Models in Animal Breeding.* University of Guelph, Ontario, Canada

Henderson, I. R. and S. E. Jacobsen, 2007. Epigenetic inheritance in plants. Nature, 447(7143): 418–424

Illingworth, R. S. and A. P. Bird, 2009. CpG islands – "A rough guide". FEBS Lett., 583(11): 1713–1720

Jeltsch, A., 2002. Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. ChemBioChem, 3(4): 274–293

Ji, H., L. I. Ehrlich, J. Seita, *et al.*, 2010. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature, 467(7313): 338–342

Jiang, Y. H., J. Bressler, and A. L. Beaudet, 2004. Epigenetics and human disease. Annu. Rev. Genomics Hum. Genet., 5: 479–510

Johannes, F., E. Porcher, F. K. Teixeira, *et al.*, 2009. Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet., 5(6): e1000530

Jones, M. C., J. S. Marron, and S. J. Sheather, 1996. A brief survey of bandwidth selection for density estimation. J. Amer. Statist. Assoc., 91: 401–407

Jones, P. A. and S. B. Baylin, 2002. The fundamental role of epigenetic events in cancer. Nat. Rev. Genet., 3(6): 415–428

Jones, P. A. and S. B. Baylin, 2007. The epigenomics of cancer. Cell, 128(4): 683–692

Kaikkonen, M. U., M. T. Lam, and C. K. Glass, 2011. Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc. Res., 90(3): 430–440

Kaput, J. and T. W. Sneider, 1979. Methylation of somatic vs germ cell DNAs analyzed by restriction endonuclease digestions. Nucleic Acids Res., 7(8): 2303–2322

Kimeldorf, G. S. and G. Wahba, 1971. Some results on tchebycheffian spline functions. J. Math. Anal. Appl., 33: 82–95

Kouzarides, T., 2007. Chromatin modifications and their function. Cell, 128(4): 693–705

Krueger, F., B. Kreck, A. Franke, *et al.*, 2012. DNA methylome analysis using short bisulfite sequencing data. Nat. Methods, 9(2): 145–151

Lim, D. H. and E. Maher, 2010. DNA methylation: a form of epigenetic control of gene expression. The Obstetrician & Gynaecologist, 12(1): 37–42

Lister, R., R. C. O'Malley, J. Tonti-Filippini, *et al.*, 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell, 133(3): 523–536

MacArthur, D., 2008. Why do genome-wide scans fail? Genetic Future (online), http://www.wired.com/2008/09/why-do-genome-wide-scans-fail/

Meijers-Heijboer, E. J., L. A. Sandkuijl, H. G. Brunner, *et al.*, 1992. Linkage analysis with chromosome 15q11-13 markers shows genomic imprinting in familial Angelman syndrome. J. Med. Genet., 29(12): 853–857

Meissner, A., A. Gnirke, G. W. Bell, *et al.*, 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res., 33(18): 5868–5877

Meissner, A., T. S. Mikkelsen, H. Gu, *et al.*, 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature, 454(7205): 766–770

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4): 1819–1829

Morota, G., P. Boddhireddy, N. Vukasinovic, *et al.*, 2014. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. Front. Genet., 5: 56

Morota, G. and D. Gianola, 2014. Kernel-based whole-genome prediction of complex traits: a review. Front. Genet., 5: 363

Morota, G., M. Koyama, G. J. Rosa, *et al.*, 2013. Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. Genet. Sel. Evol., 45: 17

Mrode, R., 2014. *Linear Models for the Prediction of Animal Breeding Values.* CAB International, 3rd edition

Nadaraya, E. A., 1964. On estimating regression. Theory Probab. Appl., 9: 141–142

Nicholls, R. D., S. Saitoh, and B. Horsthemke, 1998. Imprinting in Prader-Willi and Angelman syndromes. Trends Genet., 14(5): 194–200

Pembrey, M., 2012. *An introduction to the Genetics and Epigenetics of Human Disease.* Progress Educational Trust, London, UK

Portela, A. and M. Esteller, 2010. Epigenetic modifications and human disease. Nat. Biotechnol., 28(10): 1057–1068

Rakyan, V. K., T. A. Down, D. J. Balding, *et al.*, 2011. Epigenome-wide association studies for common human diseases. Nat. Rev. Genet., 12(8): 529–541

Ratel, D., J. L. Ravanat, F. Berger, *et al.*, 2006. N6-methyladenine: the other methylated base of DNA. Bioessays, 28(3): 309–315

Razin, A. and H. Cedar, 1991. DNA methylation and gene expression. Microbiol. Rev., 55(3): 451–458

Reinders, J., B. B. Wulff, M. Mirouze, *et al.*, 2009. Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. Genes Dev., 23(8): 939–950

Riggs, A. D., R. A. Martienssen, and V. E. A. Russo, 1996. Introduction. In V. E. A. Russo, R. A. Martienssen, and A. D. Riggs (Editors), *Epigenetic Mechanisms of Gene Regulation*, pp. 1–4. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Riggs, A. D. and T. N. Porter, 1996. Overview of epigenetic mechanisms. In V. E. A. Russo, R. A. Martienssen, and A. D. Riggs (Editors), *Epigenetic Mechanisms of Gene Regulation*, pp. 29–45. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Rivera, R. M. and L. B. Bennett, 2010. Epigenetics in humans: an overview. Curr. Opin. Endocrinol. Diabetes Obes., 17(6): 493–499

Robertson, K. D., 2002. DNA methylation and chromatin – unraveling the tangled web. Oncogene, 21(35): 5361–5379

Ruthenburg, A. J., H. Li, D. J. Patel, *et al.*, 2007. Multivalent engagement of chromatin modifications by linked binding modules. Nat. Rev. Mol. Cell Biol., 8(12): 983–994

Silverman, B., 1986. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London, UK

Tollefsbol, T. (Editor), 2012. *Epigenetics in Human Disease.* Academic Press, MA, USA

Tusell, L., P. Perez-Rodriguez, S. Forni, *et al.*, 2014. Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. J. Anim. Breed. Genet., 131(2): 105–115

van der Graaf, A., R. Wardenaar, D. A. Neumann, *et al.*, 2015. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proc. Natl. Acad. Sci. U.S.A., p. 10.1073/pnas.1424254112 (preprint)

VanRaden, P. M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci., 91(11): 4414–4423

Vanyushin, B. F., 2006. DNA methylation in plants. Curr. Top. Microbiol. Immunol., 301: 67–122

Vázquez, A. I., H. W. Wiener, S. Shrestha, *et al.*, 2014. Integration of multi-layer omic data for prediction of disease risk in humans. In *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production.* Vancouver, Canada

Waalwijk, C. and R. Flavell, 1978. DNA methylation at a CCGG sequence in the large intron of the rabbit β-globin gene: tissue-specific variations. Nucleic Acids Res., 5(12): 4631–4642

Wahba, G., 1990. *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia

Wahba, G., 1999. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. In B. Schölkopf, C. J. C. Burges, and A. J. Smola (Editors), *Advances in Kernel Methods: Support Vector Learning*, pp. 68–88. The MIT Press

Wahba, G., 2002. Soft and hard classification by reproducing kernel Hilbert space methods. Proc. Natl. Acad. Sci. U.S.A., 99: 16524–16530

Watson, G. S., 1964. Smooth regression analysis. Sankhyā: The Indian J. of Statist. Ser. A, 26: 359–372

Weber, M., J. J. Davies, D. Wittig, *et al.*, 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet., 37(8): 853–862

Zhou, H., H. Hu, and M. Lai, 2010. Non-coding RNAs and their epigenetic regulatory mechanisms. Biol. Cell, 102(12): 645–655

Chapter 7

Future Perspectives and Concluding Remarks

Biological researchers have focused on genetic variation usable for animal and plant production, as well on its impact in epidemiology. In agriculture, genetic variation is the basis of artificial selection for production improvement. In human genetics, on the other hand, study of genetic variation is helpful for finding causes of various diseases, which could be useful for developing effective treatments in the era of "personalized medicine". Variation has been investigated at the phenotypic level and at the DNA level. In recent years, epigenetics has become an important topic in such research area since it explains variation and gene expression regulations that do not reside in the DNA sequence. Therefore, associating phenotypic variation with variation conveyed by epigenomic information is important.

7.1 Contribution of Epigenetics to Quantitative Trait Variation

In livestock species, imprinting has been confirmed to be associated with various economically important traits. One example is the Callypyge locus involved in muscular hindquarter phenotype

in sheep when the *CLPG* mutation is inherited from the sire (Georges *et al.*, 2003). In breeding programs where crossing inbred lines is needed, finding contributions of genomic imprinting to genetic variation could be even more important. For example, in production of hybrids, a breeder may consider fixing a gene in the paternal line instead of the maternal line if this gene is maternally imprinted and has a positive impact on phenotypic performance. Given this potential importance of genomic imprinting in agriculture, intensive investigations of imprinted genes on farm animals and crops have been performed in recent years, and many studies suggested potential from considering genomic imprinting in agriculture (e.g., Gehring, 2013; Tian, 2014; Magee *et al.*, 2014).

Usually, an assessment of this type requires identification of imprinted genes and of their biological function in advance. One strategy is through candidate gene approaches by checking the expression or methylation profile of a genomic region whose homologous location has been identified as imprinted in other species. However, negative results can be found due to species-specific imprinting. On the other hand, a candidate region can be found through phenotype-genotype associations by considering alternative genotype phases. This initial check is not too involved, given the availability of high density SNP markers and genotype phasing algorithms. As indicated in Chapter 4, when a complex trait is subject to imprinting, this approach may detect genomic regions that are possibly imprinted. This type of study helps understanding the underlying architecture of the trait of interest, and informs on the variation contributed by imprinting. Given results from a preliminary check for imprinting, one should be able to determine downstream actions on breeding programs accordingly. Once imprinting is detected, it is possible to integrate it in genetic merit prediction of selection candidates, because imprinting is heritable across generations and, hence, may have an impact on the next generation's performance when considered suitably (Chapter 5).

Chapters 4 and 5 suggested an alternative approach for the integration of a particular epigenetic phenomenon into agricultural production without extra experimental efforts at the molecular level, as long as SNP data for genetic analysis is available. However, DNA polymorphisms information used in these studies may not be as reliable as expected when examining associations between

phenotypic variation and epigenetic mechanisms, since epigenetic modification does not change the underlying DNA sequence. Therefore, a whole-genome prediction study was conducted using MeDIP data at the epigenetic level, as described in Chapter 6. Results suggested that epigenetic information performed well as the only variation proxy when DNA polymorphisms had a negligible impact on phenotype, indicating a useful application of constructing phenotypic-epigenetic variation association studies when the impact of epigenetic modification on complex traits is of interest.

In summary, results from Chapters 4 – 6 suggested that epigenetic polymorphisms may give a detectable contribution to the phenotypic variation. Therefore, integration of epigenetic variation in a mixed effect model approach may assist breeding programs and improve agricultural production accordingly.

7.2 Biological Pathway Network Inference and Epigenetics

The potential importance of genomic imprinting and epigenetic regulation on agriculture was emphasized in the previous section from various standpoints. However, the examples given before have limitations in at least two respects. First, studies performed in Chapters 4 – 6 were based on phenotype-(epi-)genotype associations. In this type of study, with regression models used mainly, the tightness of the association is usually evaluated by statistics such as the coefficient of determination (R^2) of alternative models used in GWAS, or predictive correlation in prediction studies. These statistics provide guidelines for testing alternative hypotheses and, hence, this type of association-based method is useful for exploring phenotypic variation attributed to (epi-)genetic variants or environmental factors, and usually performs well in making outcome predictions. The drawbacks of an association-based study is that it can rarely provide inference on the "causal path" of various observations, which is important in biological studies. Molecular identification of imprinted genes may solve this "causality" problem to some extent. However, it cannot overcome a second limitation, that is, imprinted genes verified so far are mostly single discoveries. In other words, although

one may be able to discover a specific gene through molecular experiments and verify its biological functions, its interactions with other genes or other biological processes are largely ignored, especially in the case of imprinted gene identification. Since an organismal system usually represents an integration of thousands of biological processes, obtaining a comprehensive understanding of a gene's direct functionality and many other related effects is important.

Constructing a network of genes that represents their interactions and dependencies is a possible way of investigating biological processes at a broader scope. This type of study, known as gene regulatory network (GRN) inference (van Someren *et al.*, 2002), has drawn intensive interest in recent years due to the availability of high-throughput biological data. Approaches used to construct GRN models can be divided into four major categories: Boolean networks, differential equations, information theory models, and Bayesian networks (Hecker *et al.*, 2009; Imoto *et al.*, 2014), with the latter being most widely used in practice (Bornholdt, 2008; Gardner *et al.*, 2003; Stuart *et al.*, 2003; Heckerman, 2008). Many studies have reported application of GRN inference using Bayesian networks, and expression data was used as the primary input information (e.g., Friedman *et al.*, 2000; Perrin *et al.*, 2003; Ebrahimi, 2014). One advantage of Bayesian networks is their flexibility of combining multiple sources of information as prior knowledge (e.g., Werhli and Husmeier, 2007); hence, one would expect that the quality of network construction could be improved if different types of "omics" data is integrated (Hecker *et al.*, 2009). It has been reported that epigenetic information can be used as a Bayesian prior in GRN construction (Zheng *et al.*, 2011; Chen *et al.*, 2013). This type of integration has introduced an epigenomic layer into the genomic layers and, hence, improved the accuracy of network inference. According to Zheng *et al.* (2011), the sensitivity of a GRN model was improved by more than 10% with the inclusion of epigenetic prior (histone modification data), and a lower false positive rate was observed when epigenetic data was used as the sole input information, relating to a model using expression data only. Given the fact that DNA methylation is epigenetically more stable than histone modification, one may expect that methylation profiles can further improve the performance of a GRN model, comparing with the

histone modification data used in Zheng *et al.* (2011) and Chen *et al.* (2013). Besides, the two different epigenetic processes, histone modification and DNA methylation may represent different sets of the so-called "epigenetic code" (Zheng *et al.*, 2011), so it could be more informative if both sources of information are integrated into GRN inference using Bayesian networks.

A well-modeled gene-gene network may provide a more comprehensive understanding of a certain biological process and, potentially, it could be further complemented by trait-trait network inferences that explore dependencies among different traits of interest using structural equation models (e.g., Rosa *et al.*, 2011; Peñagaricano *et al.*, 2013; Valente and Rosa, 2013), a method similar to Bayesian networks (Pearl, 2012). These studies have investigated biological pathways from a different perspective. Further, network construction involving genes or QTL and complex traits has been reported as well (e.g., Schadt *et al.*, 2005; Li *et al.*, 2006), which was driven by the motivation of constructing gene-trait networks that describe causal effects of certain genes on traits of interest as well as regulatory pathways within different genes and interactive relationships within various phenotypes, an ultimate goal of biological studies. With its application in gene-gene network inference studies, it is anticipated that epigenetic information could be useful in trait-trait or gene-trait network construction as well.

7.3 Conclusion

Epigenetic regulation brings new insights into biological studies, including applications in agriculture. Among all epigenetic mechanisms, genomic imprinting is probably the most important one. Although the number of identified imprinted genes is relatively small and so far only few have been found to make significant contributions to economically important traits in agriculture, imprinting might be much more widely prevalent than previously thought (Sha, 2008). Besides, many recent studies have suggested that the silent allele is not always silent and that many imprinted genes do not show 100% monoallelic expression (Danilevskaya *et al.*, 2003; Gutierrez-Marcos *et al.*, 2004; Holl

et al., 2004; Köhler *et al.*, 2005; Khatib, 2007). The meaning of this partial imprinting is yet unclear (Gehring, 2013), but it may also contribute to variation of complex traits. Given the regulatory role of epigenetic mechanisms without changes of the DNA sequence, the integration of epigenetic information may be helpful for improving agricultural production and for attaining a better understanding of biology systems via pathway analysis. These aspects deserve more investigation in the future.

References

Bornholdt, S., 2008. Boolean network models of cellular regulation: prospects and limitations. J. R. Soc. Interface, 5 (Suppl. 1): 85–94

Chen, H., D. A. K. Maduranga, P. A. Mundra, *et al.*, 2013. Integration epigenetic prior in dynamic Bayesian network for gene regulatory network inference. In *Proceedings of the 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (IEEE CIBCB 2013)*, pp. 76–82. Singapore

Danilevskaya, O. N., P. Hermon, S. Hantke, *et al.*, 2003. Duplicated *fie* genes in maize: expression pattern and imprinting suggest distinct functions. Plant Cell, 15(2): 425–438

Ebrahimi, A., 2014. *Inferring Gene regulatory network based on Bayesian network: Gene expression.* LAP LAMBERT Academic Publishing

Friedman, N., M. Linial, I. Nachman, *et al.*, 2000. Using Bayesian networks to analyze expression data. J. Comput. Biol., 7(3-4): 601–620

Gardner, T. S., D. di Bernardo, D. Lorenz, *et al.*, 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. Science, 301(5629): 102–105

Gehring, M., 2013. Genomic imprinting: insights from plants. Annu. Rev. Genet., 47: 187–208

Georges, M., C. Charlier, and N. Cockett, 2003. The callipyge locus: evidence for the *trans* interaction of reciprocally imprinted genes. Trends Genet., 19(5): 248–252

Gutierrez-Marcos, J. F., L. M. Costa, C. Biderre-Petit, *et al.*, 2004. *maternally expressed gene 1* is a novel maize endosperm transfer cell-specific gene with a maternal parent-of-origin pattern of expression. Plant Cell, 16(5): 1288–1301

Hecker, M., S. Lambeck, S. Toepfer, *et al.*, 2009. Gene regulatory network inference: data integration in dynamic models-a review. BioSystems, 96(1): 86–103

Heckerman, D., 2008. A tutorial on learning with Bayesian networks. In D. E. Holmes and L. C.

Jain (Editors), *Innovations in Bayesian Networks: Theory and Applications*, pp. 33–82. Springer

Holl, J. W., J. P. Cassady, D. Pomp, *et al.*, 2004. A genome scan for quantitative trait loci and imprinted regions affecting reproduction in pigs. J. Anim. Sci., 82(12): 3421–3429

Imoto, S., H. Matsuno, and S. Miyano, 2014. Gene networks: estimation, modeling, and simulation. In A. Kriete and R. Eils (Editors), *Computational Systems Biology: From Molecular Mechanisms to Disease*, pp. 89–112. Academic Press, Oxford, 2nd edition

Khatib, H., 2007. Is it genomic imprinting or preferential expression? Bioessays, 29(10): 1022–1028

Köhler, C., D. R. Page, V. Gagliardini, *et al.*, 2005. The *Arabidopsis thaliana* MEDEA Polycomb group protein controls expression of PHERES1 by parental imprinting. Nat. Genet., 37(1): 28–30

Li, R., S. W. Tsaih, K. Shockley, *et al.*, 2006. Structural model analysis of multiple quantitative traits. PLoS Genet., 2(7): e114

Magee, D. A., C. Spillane, E. W. Berkowicz, *et al.*, 2014. Imprinted loci in domestic livestock species as epigenomic targets for artificial selection of complex traits. Anim. Genet., 45 (Suppl. 1): 25–39

Peñagaricano, F., K. A. Weigel, G. J. M. Rosa, *et al.*, 2013. Inferring quantitative trait pathways associated with bull fertility from a genome-wide association study. Front. Genet., 3: 307

Pearl, J., 2012. On Structural Equations versus Causal Bayes Networks. (online), http://www.mii.ucla.edu/causality/?p=571

Perrin, B. E., L. Ralaivola, A. Mazurie, *et al.*, 2003. Gene networks inference using dynamic Bayesian networks. Bioinformatics, 19 (Suppl. 2): ii138–ii148

Rosa, G. J. M., B. D. Valente, G. de los Campos, *et al.*, 2011. Inferring causal phenotype networks using structural equation models. Genet. Sel. Evol., 43: 6

Schadt, E. E., J. Lamb, X. Yang, *et al.*, 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet., 37(7): 710–717

Sha, K., 2008. A mechanistic view of genomic imprinting. Annu. Rev. Genomics Hum. Genet., 9: 197–216

Stuart, J. M., E. Segal, D. Koller, *et al.*, 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science, 302(5643): 249–255

Tian, X. C., 2014. Genomic imprinting in farm animals. Annu. Rev. Anim. Biosci., 2: 23–40

Valente, B. D. and G. J. M. Rosa, 2013. Mixed effects structural equation models and phenotypic causal networks. Methods Mol. Biol., 1019: 449–464

van Someren, E. P., L. F. Wessels, E. Backer, *et al.*, 2002. Genetic network modeling. Pharmacogenomics, 3(4): 507–525

Werhli, A. V. and D. Husmeier, 2007. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. Stat. Appl. Genet. Mol. Biol., 6(1): Article 15

Zheng, J., I. Chaturvedi, and J. C. Rajapakse, 2011. Integration of Epigenetic Data in Bayesian Network Modeling of Gene Regulatory Network. In M. Loog, L. Wessels, M. J. T. Reinders, and D. de Ridder (Editors), *Pattern Recognition in Bioinformatics: 6th IAPR International Conference*, pp. 87–96. Delft, The Netherlands