ACADEMY OF WALLONIA-EUROPE
UNIVERSITY OF LIÈGE
FACULTY OF VETERINARY MEDECINE
DEPARTMENT OF ANIMAL PRODUCTION
UNIT OF ANIMAL GENOMICS

POSITIONAL IDENTIFICATION OF A REGULATORY MUTATION IN THE PORCINE *IGF2* GENE INFLUENCING MUSCLE MASS & FAT DEPOSITION

CLONAGE POSITIONEL D'UNE MUTATION REGULATRICE DANS LE GÈNE *IGF2* PORCIN INFLUENCANT LA MASSE MUSCULAIRE ET LA QUANTITÉ DE GRAS

Minh NGUYEN

THESIS PRESENTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHYLOSOPHY IN VETERINARY SCIENCE ORIENTATION: ANIMAL PRODUCTION

ACADEMIC YEAR: 2009-2010

ACKNOWLEGEMENTS

It would not have been possible to accomplish this Ph.D. thesis without the help and contributions from a number of people throughout the course of my Doctorate; unfortunately it will be possible to acknowledge only a few here.

Above all, Prof. Michel Georges, promoter of my thesis: I would like to express my deepest gratitude to you for giving me the rare opportunity to work in your cutting-edged laboratory and on this fascinating research project. Word alone can not express my appreciation for your supervising, motivating, and especially "revising" my manuscripts critically. From you not only did I discover the wonders of complex traits and positional cloning, genetics and epigenetics but also learnt to become a scientist. Michel, it is my real honor to know you.

Mrs. Bernadette Marcq: I am grateful to you for helping me so much on administrative documents each year and for "a few" years long! I recall your cheerfulness on the first day I arrived at the Gare de Guillemins, all the drinks, parties and the amazing Fete de Noel that I had never experienced before. Bernadette, I should have done better, shouldn't I?

Dr. Carole Charlier, my co-promoter: your straightforward responses to and fruitful discussions on all my questions, from bisulphite sequencing to advice on thesis reading and writing, casual talks, smiling, and "teasing", all were appreciated.

Members of my thesis committee and jury: Prof. Vincent BOURS, Prof. Joseph MARTIAL, Prof. Juliette RIQUET, Prof. Luc PEELMAN, Dr. Jean-Francoise CABARAUX, Prof. Antoine CLINQUART, Prof. Fédéric FARNIR, Dr. Charles MICHAUX, and Dr. Martine LAITAT. I would like to thank you for your time and comments on thesis writing and public defense.

This work has been rooted from and contributed to by a number of groups. My special thanks to Dr. Carine Nezer, Mrs. Laurence Moreau, Mr. Benoît Brouwers, and Mrs. Catherine Collette - the pig QTL group – University of Liège for their investigation of the QTL that provided the basis for my Ph.D. thesis; to Prof. Leif Andersson, Dr. Valerie Amarger, Dr. Martin Braunschweig, and Dr. Anne-Sophie Vanlaere – the genetics group in the Swedish University of Agricultural Sciences; to Prof. Nadine Buys – Seghers Genetics for all their support and collaborations.

Mrs. Latifa Karim: in particular, not only do I want to thank you for all your help with sequencing on the ABI 377, interrupting you any time with requests, but also for your lively charming talks. You and your family made my stay more cheerful in Liège.

Dr. Dominique Poncelet, Prof. Luc Grobet, and Dr. Dimitri Pirotin: Thank you all for sharing your knowledge, consistent explanation, suggestions, and fruitful discussions. From you I have learnt real molecular biology. You are smart and sincere, aren't you?

I particularly thank all people who have helped me along my Ph.D. program including those in the Unit of Animal Genomics, Bioinformatics - Computer groups, Porcine Clinics, Collège de Doctorat de la Faculté de Médecine Vétérinaire, administrative departments, and libraries.

Friends: thank you all Vietnamese, Belgian, and international friends both in Brussels and Liège with whom I frequently met and shared time. Thanks God for bringing us together!

My Mom and my grand family: Món quà nhỏ bé này con dành tặng Mẹ và cả nhà, cho sự hy sinh và tình thương yêu vô bờ bến suốt cuộc đời của Mẹ dành cho chúng con.

My wife: Thank you for your love and taking the risk of marrying me, for your daily support during the last but critical period of my Ph.D.

This work was supported by grants from Belgian Ministères des Classes Moyennes et de l'Agriculture and Gentec. I would like to acknowledge their continuous support throughout my study.

I would like to thank people in National Institute of Veterinary Research, Ministry of Agriculture and Rural Development, Ministry of Education and Training in Hanoi – Vietnam, Vietnamese embassy in Brussels for their support during my study in abroad.

Undoubtedly, this Ph.D. thesis was accomplished by the combined efforts of many people. Though their names are mentioned herein or not, their support and contributions are engraved in my mind now and then.

Minh Nguyen

Liège, Belgium

September 2009

SUMMARY

Recent advances in genomics now allow for the identification of the genes and mutations that underlie the heritability of agronomically important traits in livestock. The corresponding genes are said to map to Quantitative Trait Loci (QTL), and the mutations referred to as Quantitative Trait Nucleotides (QTN). The most commonly used approach relies on positional cloning which typically proceeds in three steps: QTL mapping by linkage analysis, QTL fine-mapping by linkage disequilibrium or association analysis, and QTN identification combining haplotype analysis and functional assays. Knowledge of QTL and QTN provides insights into the genetic architecture of complex traits and physiology of production traits, and opens novel possibilities for enhanced selection referred to as Marker Assisted Selection (MAS).

This thesis is devoted to QTN identification of a QTL that was previously mapped to pig chromosome 2 and fine-mapped to a 250 Kb segment encompassing the imprinted *IGF2* gene. The QTL was shown to have a major post-natal effect on muscle mass and fat deposition, and to be subject to parental imprinting as only the padumnal chromosome affects the phenotype.

To identify the QTN we have first generated 32 Kb and 56 Kb of finished porcine sequence encompassing the *IGF2* and *H19* genes, respectively. The corresponding sequences were annotated including definition of gene models, identification of interspersed repeats and determination of 97 sequence elements that are highly conserved between pig, human and mouse.

We have then resequenced 28 Kb encompassing the *IGF2* gene for 15 boar chromosomes for which the QTL genotype had been determined by progeny-testing or Marker Assisted Segregation Analysis (MASA). This revealed 258 polymorphisms of which only one (Int3-3072G>A) cosegregated perfectly with QTL genotype. The corresponding single nucleotide polymorphism (SNP) is a G to A transition affecting one of the highly conserved sequence elements located just downstream of differentially methylated region 1 in intron 3. We have demonstrated that the Int3-3072 A allele associated with increased muscle mass is also associated with increased *IGF2* mRNA levels in post-natal striated muscle (but not in pre-natal muscle nor in pre- and post-natal liver). However, the Int3-3072G>A SNP does not alter imprinting nor allele-specific methylation. Using a luciferase reporter assay, we then demonstrated that the Int3-3072 A allele reduces the *cis* activity of a silencer element, and using an electrophoretic mobility shift assay (EMSA), that it abrogates binding of a nuclear factor assumed to be a *trans*-acting silencing factor. Taken together both genetic and functional evidence strongly support the conclusion that the Int3-3072G>A SNP is the causative SNP.

The thesis is concluded by a discussion that (i) highlights the factors that make domestic animals a unique resource for the molecular dissection of complex phenotypes, (ii) comments the Asian origin of the Int3-3072A allele associated with increased muscle mass, (iii) describes recent advances in characterizing the trans-acting silencing factor binding to the Int3-3072G allele, (iv) pinpoints statistical issues related to the detection of imprinted QTL, (v) reports on the utility of the Int3-3072G>A SNP for MAS applied to pig breeding, and (vi) makes projections on how latest progress in genome analysis will affect positional identification of QTN in the near future.

RÉSUMÉ

Grâce aux progrès récents en génomique, il est maintenant possible d'identifier les gènes et mutations qui sous-tendent l'héritabilité des caractères de production chez les animaux de rente. Ces gènes se localisent au niveau de Loci de Traits Quantitatifs (QTL), et les mutations correspondantes sont qualifiées de Nucléotides de Traits Quantitatifs (QTN). La démarche expérimentale la plus couramment utilisée est le clonage positionnel. Celui-ci comprend trois étapes: cartographie de QTL par analyses de liaison génétique, cartographie fine de QTL par études d'association exploitant le déséquilibre de liaison, et identification de QTN par combinaison d'analyses haplotypiques et fonctionnelles. L'identification de QTL et QTN non seulement révèle l'architecture génétique des phénotypes complexes que sont les caractères de production, ainsi que les rouages moléculaires qui les sous-tendent, mais ouvre également des possibilités nouvelles de sélection plus performante dite Assistée par Marqueurs (MAS).

Cette thèse est consacrée à l'identification d'un QTN correspondant à un QTL d'abord localisé sur le chromosome 2 du porc, et ensuite cartographié finement dans un segment chromosomique de 250 Kb comprenant le gène *IGF2*. Le QTL en question a un effet post-natal majeur sur la croissance musculaire et le dépôt graisseux. Il est soumis à l'empreinte parentale, l'allèle paternel étant le seul à influencer le phénotype. Afin d'identifier le QTN, nous avons tout d'abord généré 32 Kb et 56 Kb de séquences finies, comprenant respectivement les gènes *IGF2* et *H19*. Les séquences correspondantes ont été annotées bioinformatiquement, y compris la définition de modèles géniques, l'identification de séquences répétées dispersées, ainsi que de 97 éléments de séquence fortement conservés chez le porc, l'homme et la souris.

Nous avons ensuite re-séquencé 28 Kb chevauchant le gène *IGF2* pour 15 chromosomes dont le génotype au niveau du QTL fut préalablement déterminé par testage de descendance ou Ségrégation Assistée par Marqueurs. Cet exercice a révélé 258 polymorphismes dont un seulement (Int3-3072G>A) correspondait parfaitement aux génotypes QTL. Ce polymorphisme est une transition G à A affectant un des 97 éléments hautement conservés, situé juste en aval de la région différentiellement méthylée (DMR1) dans l'intron 3. Nous avons ensuite démontré que l'allèle Int3-3072 A, associé à une augmentation de la masse musculaire, est également associé à une augmentation des taux d'ARNm *IGF2* dans le muscle strié post-natal (mais non dans le muscle strié pré-natal, ni dans le foie pré- et post-natal). Par contre, le polymorphisme Int3-3072G>A n'affecte ni état d'empreinte ni de méthylation allèle-spécifique du gène. Nous avons ensuite démontré à l'aide d'un test rapporteur de type luciférase que l'allèle Int3-3072 A réduit l'activité d'un élément silenceur agissant en *cis*, et à l'aide d'un test de type EMSA qu'il empêche la liaison d'un facteur nucléaire. Conjointement, ces données génétiques et fonctionnelles démontrent que le polymorphisme Int3-3072G>A SNP correspond bien au QTN.

Nous concluons la thèse par une discussion dans laquelle nous (i) démontrons pourquoi les animaux domestiques offrent des possibilités uniques pour la dissection moléculaire de phénotypes complexes, (ii) commentons l'origine asiatique de l'allèle Int3-3072A associé à une augmentation du développement musculaire, (iii) décrivons les progrès récent dans l'identification du facteur nucléaire reconnaissant spécifiquement l'allèle Int3-3072G, (iv) attirons l'attention sur les artéfacts statistiques associés à la détection de QTL soumis à l'empreinte, et (v) discutons l'impact de nouvelles technologies génomiques sur le clonage positionnel de gènes.

TABLE OF CONTENTS

ABBREVIATIONS

A:	Adenosine
Ab:	Antibody
AI:	Artificial Selection
BAC:	Bacterial Artificial Chromosome
bp:	base pair
BSA:	Bovine Serum Albumin
C:	Cytosine
C1/2/3:	Complex 1/2/3
C2C12:	Mouse myoblast cell line from C3H strain
cDNA:	Complementary DNA
Chr:	Chromosome
CLPG:	*Callipyge*
cM:	CentiMorgan
CpG:	Cytosine-Guanosine dinucleotides
CS:	Conserved Sequence
CTCF:	CCCTC-binding factor
DBA:	DNA Block Aligner
DMR:	Differential Methylation Region
DNA:	Deoxyribonucleic Acid
DNMT:	*DNA methyltransferases*
EMSA:	Electrophoretic Mobility Shift Assay
EPD:	Eukaryotic Promoter Database
EST:	Expressed Sequence Tag
EWB:	European Wild Boar
F0, F1, F2, F3:	Generation 0, 1, 2, 3
G:	Guanosine
GAPDH:	*GlycerAldehyde 3-Phosphate DeHydrogenase* (house keeping gene)
GCF2:	GC binding factor
GS:	Genomics Selection
H/M:	Human/Mouse
H:	Hampshire
HEK293:	Human embryonic kidney fibroblast cell line
HepG2:	Hepatocytes G2
HSA11:	Homo Sapiens (Human) Chromosome 11
ICR:	Imprinting Control Region
IDDM:	Insulin-Dependent Diabetes Mellitus
IGF/Igf:	Insulin-like Growth Factor
IGFBP:	Insulin-like Growth Factor Binding Protein
IGF-R:	Insulin-like Growth Factor Receptor
Ihit1:	*IGF2-H19* Interval Transcript 1
INS:	*Insulin*
JWB:	Japanese Wild Boar
Kb:	Kilo base
KDa:	Kilo Dalton
Lit1:	Long QT Intronic Transcript 1 (imprinted antisense RNA)
LOI:	Loss of Imprinting
LR:	Linear Regression
LW:	Large White

MAS:	Marker-Assisted Selection
MASA:	Marker-Assisted Segregation Analysis
Mash2:	Achaete-Scute complex-like protein 2
Mat/mat:	Maternal
Mb:	Mega base
M:	Meishan
MIR:	Mammalian Interspersed Repeat
mRNA:	Messenger Ribonucleic Acid
MAR:	Matrix Attachment Region
MSTN:	Myostatin
MYOD:	Myoblast Determination Protein
Nctc1:	Non-coding transcript 1
NE:	Nuclear Extracts
P/H:	Pig/Human
P/M:	Pig/Mouse
P1/2/3/4:	Promoter 1/2/3/4
P208:	Piétrain 208
P3–LUC:	Promoter 3 fused with Luciferase gene
Pat/pat:	Paternal
PCR:	Polymerase Chain Reaction
PFGE:	Pulsed Field Gel Electrophoresis
Q (-haplotype):	Haplotype carrying mutation with bigger effect on phenotype
q (-haplotype):	Haplotype carrying wild type with smaller effect on phenotype
QTL:	Quantitative Trait Loci
QTN:	Quantitative Trait Nucleotide
RACE:	Rapid Amplification of cDNA Ends
Rhit1:	Alias of *Nctc1*
RIA:	Radio Immuno Assay
RNA:	Ribonucleic Acid
RT-PCR:	Reverse Transcriptase - Polymerase Chain Reaction
SINE:	Short Interspersed Repeats
SM:	Skeletal Muscle
SNP:	Single Nucleotide Polymorphism
SSC2:	Sus Scrofa (pig) Chromosome 2
SW-C/R/T:	Names of microsatellite genetic markers
T:	Thymine
TH:	*Tyrosine Hydroxylase*
TK:	*Heterologous Herpes Thymidine Kinase*
UTR:	Untranslated Region
VNTR:	Variable Number Tandem Repeat
YBP:	Year before Presence
Yr/yr:	Year

CHAPTER 1

GENERAL INTRODUCTION

The ultimate aim of livestock production science is to improve animal performances for economically important traits in order to better meet the demands of the consumers. Livestock productivity has improved dramatically over the last fifty years, and this is to a large extent due to enhancement of genetic merit of the herds. Most economically important traits in livestock are complex phenotypes, i.e. they are influenced by environmental factors and multiple "polygenes" mapping to Quantitative Trait Loci or QTL. Traditional selection has thus very effectively, albeit blindly, let to changes of the allelic composition of the herds at these polygenes. Since the 1980-ies it has become feasible to map QTL and identify polygenes that contribute to the genetic variation for economically important traits. This might lead to even more effective selection methods, referred to as "Marker Assisted Selection" or MAS. This thesis deals with the identification and use of polygenes influencing carcass composition in the pig.

1.1. Complex traits:

1.1.1. Definition (Lynch and Walsh, 1998):

The vast majority of economically important traits in agriculture are complex traits that are influenced by environmental factors and multiple segregating genes. Most of these, such as milk yield and composition, are measured on a continuous scale and are therefore referred to as quantitative traits. The typical normal distribution observed for quantitative traits can be explained by the mainly simple addition of the modest allele substitution effects at individual polygenes. Contributing polygenes map to loci called QTL. QTL are typically identified by positional cloning involving QTL mapping.

1.1.2. Heritability (Lynch and Walsh, 1998):

Stating that complex traits are influenced both by genetic and non-genetic factors, implies that the phenotypic variance observed for a trait in a given population can be partitioned in a genetic (G) and a non-genetic (E = environment) variance component:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 \qquad (1)$$

The proportion of the phenotypic variance that is genetic in nature corresponds to the broad-sense heritability of the trait, H^2.

The genetic variance can itself be partitioned in an additive (σ_A^2) and a non-additive (σ_{NA}^2) component. The latter results from dominance effects within loci and epistatic effects between

loci. The proportion of the phenotypic variance corresponding to σ_A^2 is referred to as the narrow-sense heritability h^2. Proper estimation of h^2 is essential in animal breeding as the response to selection is a function of h^2. h^2 is typically estimated using variance component methods, comparing the observed and predicted (under an additive "animal" model) phenotypic covariance between relatives as a function of the coefficient of kinship.
Narrow-sense heritability estimates for body composition traits in swine range from 0.3 to 0.57 for killing out percentage and carcass length, respectively (table 1).

Table 1: Average values of heritability (h^2) for body composition traits (adapted from Rothschild and Ruvinsky, 1998).

Traits	Stewart and Schinckel (1989)	Ducos (1994)
Ultrasonic backfat thickness	0.41	0.45
Fat depth over the 10th rib	0.52	-
Loin muscle area	0.47	0.48
Lean percentage	0.48	0.54
Killing out percentage	0.30	0.36
Carcass length	0.56	0.57

1.2. Positional cloning:

Although traditional "mass" selection has proven remarkably effective even in the absence of any molecular knowledge of the genes that it acts upon, it is generally assumed that knowing the corresponding genes would not only be of major fundamental interest but would also allow the implementation of even more effective selection schemes referred to as MAS. Two strategies are typically used for the identification of genes underlying complex trait: the "physiological" candidate gene approach, and positional cloning. Positional cloning is "generic" in nature: it does not require any prior knowledge about gene function, which is still restricted to a minority of protein encoding genes. Thanks to the rapid development of genome-wide microsatellite and SNP-based maps, positional cloning has become the method of choice. It was selected for the research described in this thesis. Positional cloning typically

proceeds in three steps: QTL mapping, QTL fine-mapping and QTN (Quantitative Trait Nucleotide) identification.

1.2.1. QTL mapping:

Methods for the mapping of QTL are numerous and constantly evolving. Herein, I will briefly describe the basic principles of QTL interval mapping in line-crosses. The QTL that is the topic of this thesis was indeed mapped using this approach.

To map the QTL that underlie the phenotypic difference observed between breeds or lines, one typically first generates an intercross or backcross. This is achieved by mating the "parental" lines to obtain a hybrid "F1" generation which is then either mated to one or both parental lines to yield a backcross generation, or intercrossed to yield an "F2" generation. The QTL studied in this thesis was mapped in a Piétrain x Large White F2 pedigree.

Phenotypes of interest are then recorded for all animals in the pedigree, or at least for the "informative" F2 generation. In this study, collected phenotypes pertained to growth and carcass composition.

Subsequently, all animals in the pedigree are genotyped for a battery of genetic markers spanning the genome. The markers used for most QTL mapping efforts performed to date are sets of ~150 highly informative microsatellites covering the genome with average spacing of 20 centimorgans.

Finally, the presence of QTL in each of the intervals between adjacent markers is tested by comparing the phenotypic means of F2 offspring sorted in the three possible categories of genotypes (referred to herein as RR, BB and RB, assuming that the parental lines are R(ed)R(ed) and B(lue)B(lue), respectively (Fig. 1). Comparison of means can be conducted using a variety of methods, but the most popular one is "linear regression" according to Haley and Knott (1994). This approach indeed accounts effectively for the fact that the genotype of an F2 individual in an interval of interest is not known with certainty, however, that genotype probabilities (P_{RR}, P_{BB}, P_{RB}) can be estimated from flanking marker data. One can show that the regression coefficient of phenotype on (P_{RR}-P_{BB}) estimates a, i.e. half the difference between the phenotypic means of RR and BB F2 animals, while the regression coefficient of phenotype on P_{RB} estimates d, the deviation between the midpoint between the averages of the RR and BB phenotypes and the average phenotype of RB animals.

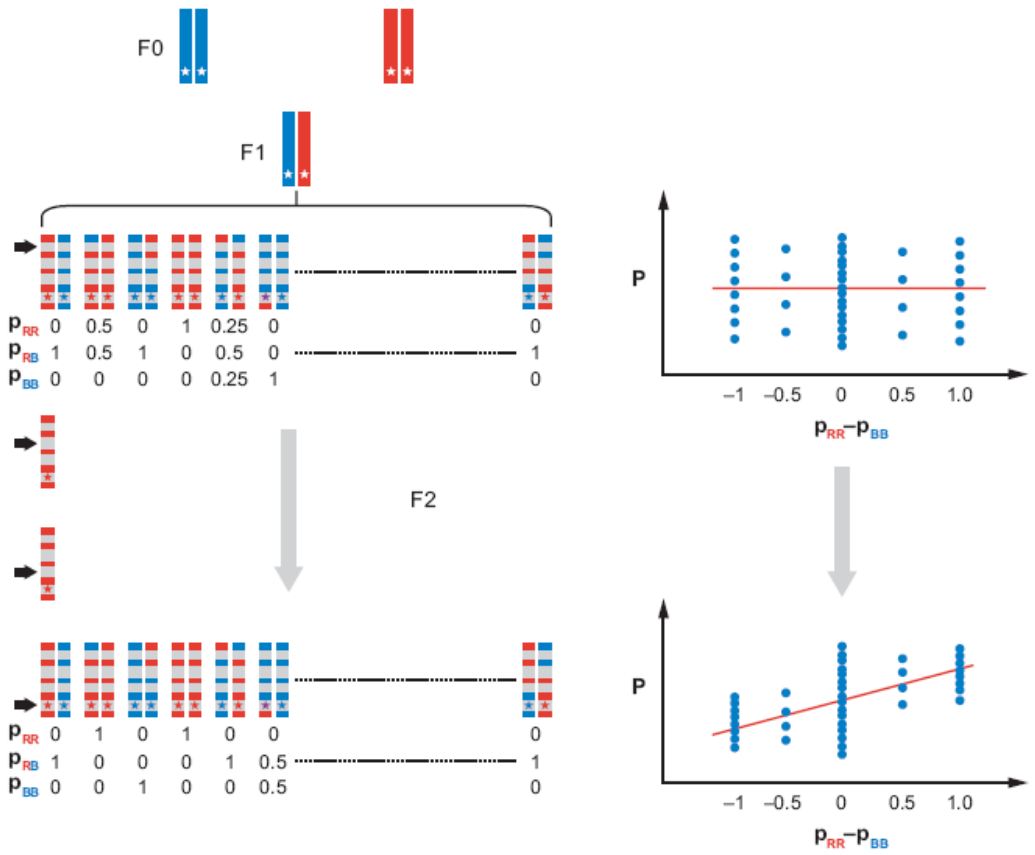


Figure 1: Principles of quantitative trait loci (QTL) interval mapping using linear regression (LR) illustrated for an F2 cross. An F2 population is generated by intercrossing "blue" and "red" parental strains differing for a phenotype of interest. The F2 population is genotyped with a battery of genetic markers covering the genome at regular intervals of ~10 centiMorgans (cM), shown as colored bars on the chromosomes of the F2 individuals. Marker intervals are "interrogated" successively (*black arrows*) for the presence of a QTL. For each interval, and for each F2 individual, one computes the probability that the individual is homozygous "red-red" (pRR), heterozygous "red-blue" (pRB), or homozygous "blue-blue" (pBB), using the observable genotypes at flanking marker loci. The additive effect of a given interval on the phenotype is estimated by regressing the phenotypes on $p_{RR}-p_{BB}$, as shown in the panels on the right. In the absence of a QTL in the tested interval (e.g., interval 1), the regression coefficient does not deviate significantly from 0. In the presence of a QTL in the corresponding interval (shown by the star in interval 4), the regression coefficient may deviate significantly from 0 (Georges, 2007).

To account for the fact that one tests multiple correlated intervals when performing a genome scan, the statistical significance of the deviations of the a and d regression coefficients from 0 is estimated by phenotype permutation (Doerge and Churchill, 1996). Confidence intervals for QTL locations are conveniently estimated by bootstrapping (Visscher et al., 1996).

A large number of QTLs has been identified in livestock. According to the current release (May 16, 2009) of the QTL databases (http://www.animalgenome.org/QTLdb/), 1,831 QTLs influencing 316 different traits were identified in the pig, 1,123 QTLs influencing 101

different traits were identified in cattle, and 657 QTLs influencing 112 different traits were identified in poultry.

1.2.2. QTL fine mapping (Georges, 2007):

Confidence intervals for the QTL are typically several tens of centimorgans long after the mapping step. In most species, this would correspond to several tens of megabases and hundreds of "positional candidate genes". To reduce the size of the confidence interval and hence the number of candidate genes, one has to refine the QTL map position. This requires reduction of the size of the marker intervals by increasing local marker density, and increasing the number of informative recombinant chromosomes in the interval, as QTL localization ultimately depends on such chromosomes only. Increasing marker density is in principle straightforward even if still time-consuming especially for species without reference genome sequence. Increasing the density of recombinational events can be achieved by generating additional F2 offspring and selecting for further analysis those having inherited a chromosome recombining in the interval of interest. This strategy however is time-consuming and expensive. When working with domestic species, it is more effective to exploit pre-existing "historical" rather than de novo generated recombinants. Fig. 2 underlies the principles of this "linkage disequilibrium"-based approach. The QTL polymorphism, that accounts for the mapping of the QTL in the first place, reflects the occurrence of a neo-mutation that created for instance a Q allele from an ancestral q allele n generations ago on a chromosome characterized by a distinct haplotype at flanking polymorphisms. Especially if the Q allele causes an advantageous phenotype, selection will have increased its frequency in subsequent generations. During its expansion in the population, Q-bearing alleles will recombine with homologous chromosomes, thereby reducing the length of the ancestral haplotype shared identical-by-descent by all Q chromosomes at the present generation. The QTL can be fine-mapped by identifying this shared haplotype whose length depends on the number of generations since the birth of the Q allele, and on the number of extant Q-bearing chromosomes available for comparison. Extant Q-bearing chromosomes can best be identified by marker assisted segregation analysis (MASA), i.e. by identifying parental chromosomes (typically the chromosomes of a sire with many phenotyped offspring) whose segregation in the offspring is associated with differences in phenotypic outcome in agreement with the effect of the studied QTL both in terms of phenotypic effect and chromosomal location.

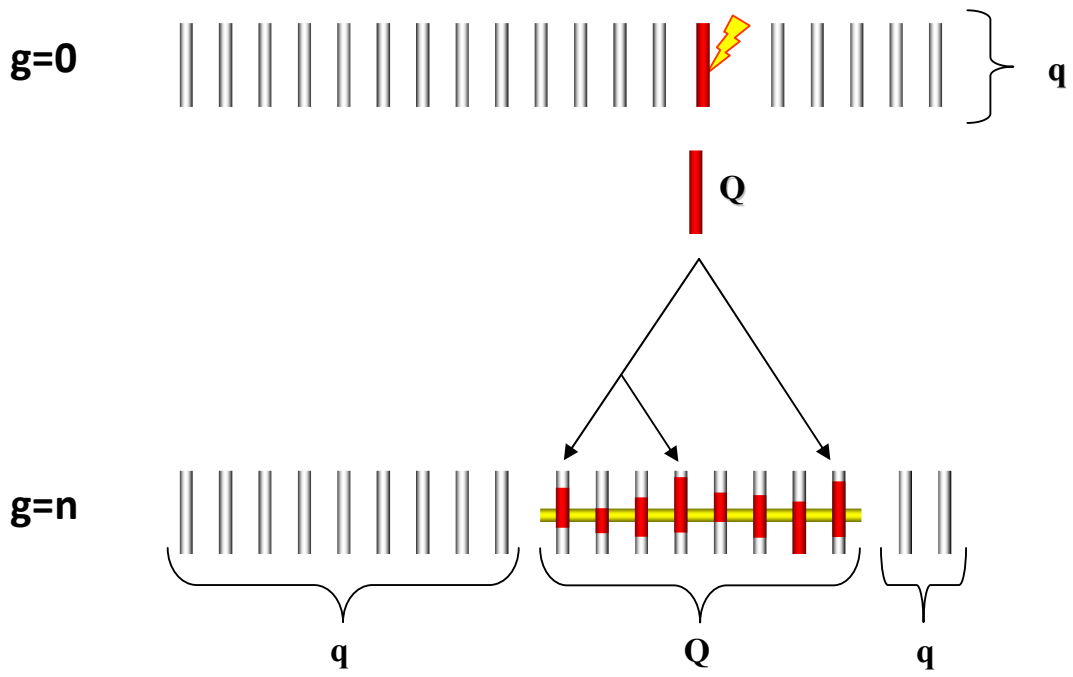


Figure 2: Principle of QTL fine mapping using "historical recombinants". A mutation transforming the wild-type q in a novel Q QTL allele is assumed to have occurred N generations ago on a chromosome labeled in red. Especially if the Q allele has a favorable effect on phenotype, the Q allele will spread in the population. As Q alleles segregate in the population, the red haplotype shared identical by descent by all of them is progressively reduced in size as a result of crossing over with non-red haplotypes. If chromosomes carrying the Q QTL allele could be identified at the present generation n, and genotyped for a high density set of markers, it should reveal a relatively small, haplotype shared by all Q chromosomes but not q chromosomes, predicted to encompass the causative QTNs.

Examples of successful QTL fine-mapping using this approach include: (i) a QTL influencing milk yield and composition on bovine chromosome 14 (Riquet et al., 1999; Farnir et al., 2002) (ii) a paternally imprinted QTL with major effect on muscle mass and fat deposition mapped to a 250 Kb chromosomal segment on SSC2 (Nezer et al., 2003) (iii) a QTL contributing to the muscular hypertrophy of Texel sheep mapped to a 2.5 cM interval (Clop et al., 2006).

1.2.3. QTN identification:

Mapping and fine-mapping relies on the use of low and high density *marker* maps respectively, i.e. the used markers serve to track haplotypes between generations but are not assumed to be the causative polymorphisms or QTN. Achieving the ultimate goal of identifying the causative polymorphisms requires (i) cataloguing of all DNA sequence polymorphisms differentiating chromosomes carrying alternative QTL alleles, (ii) genetic "sorting" of candidate polymorphisms, and (iii) evaluating the functional effects of the qualifying polymorphisms.

To be exhaustive, the first step requires complete sequencing of the confidence interval for the QTL from chromosomes with known QTL genotype. Especially since the development of high-throughput sequencing-by-synthesis procedures, this objective is becoming realistic for confidence intervals of increasing size. For the second step to be effective, as many as possible QTL-genotyped chromosomes with distinct ancestry will be sequenced. Assuming the most parsimonious model of a single causative mutation, the only DNA sequence polymorphisms that will maintain putative QTN status are those that show perfect segregation between the sets of Q and q chromosomes, i.e. all Q chromosomes should carry the same allele, distinct from that carried by all q chromosomes. Finally, one will want to demonstrate which one of the remaining candidates truly affects gene function. Reliable *ab initio* prediction of a functional effect is simple for non-sense mutation, possible for missense mutations and very difficult for any other type of polymorphisms. Genuine functional test will therefore have to be conducted including measuring the enzymatic activity of alternative allelic forms for structural mutations, or the effect on transcriptional regulation, transcript processing, and transcript translation of putative regulatory mutations. Knocking the mutation in the orthologous murine locus using gene targeting techniques and replicating the phenotypic effects would of course provide very convincing evidence for causality but is a very tedious proposition.

Very few examples of successful QTN identification have been reported to date. In livestock species, these include: (i) the *DGAT1 K232A* (Grisart et al., 2002 and 2004; Winter et al., 2002) and *ABCG2 Y581S* (Cohen-Zinder et al., 2005) mutations influencing milk yield and composition in cattle, and (ii) the MSTN 3' untranslated region g+6723(G-A) mutation influencing muscle mass in sheep (Clop et al., 2006).

1.3. Marker-Assisted Selection (MAS) and Genomic Selection (GS) (Dekkers and Hospital, 2002):

Information on QTL underlying part of the genetic variation for a quantitative trait of interest can, in principle, be used to improve response to artificial selection. Response to selection depends on four factors: (i) the accuracy of selection, (ii) the intensity of selection, (iii) the level of genetic variation, and (iv) the generation interval.

Knowledge of underlying QTL can affect each one of the four determinants of genetic response. The more the identified QTL contribute to the genetic variance of a given trait, the higher the accuracy with which one can select the individuals with highest breeding values as parents for the next generation. The relative improvement over standard selection is highest

for the traits with lowest heritability. Hence, the power of MAS is often set to be highest for low heritability traits. It has to be realized, however, that QTL are harder to detect for such traits. As a consequence, identified QTL usually explain a smaller fraction of the genetic variance and the benefits of MAS remain limited. Gains in accuracy of selection are also highest for trait that are difficult to measure on the live animal, such as carcass traits or, most strikingly, traits with sex-limited expression such as milk and egg production.

Intensity of selection has the potential to be greatly enhanced by MAS, especially when combined with reproductive technologies such as embryo transfer. Indeed, most breeding designs only allow phenotypic characterization of a limited number of animals. As an example, artificial insemination (AI) companies can only progeny- or performance-test so many bulls per season. Pre-selecting animals to test by MAS could in effect improve genetic response by increasing the intensity of selection.

The level of genetic variation could in principle be affected by MAS as well, by introgressing favorable QTL alleles discovered in a donor breed into a recipient population. The introgression process could be accelerated by counter selecting against the undesired remainder of the donor genome using genome-wide markers in a process akin to "speed-congenics" as applied to model organisms. This objective is certainly one of the possible outcomes of ongoing efforts to understand the remarkable prolificity of Chinese sows. The corresponding prolificity QTL alleles could – once identified – be introgressed in European pig breeds which have more favorable QTL alleles for carcass traits.

MAS also has the potential to have a major impact on generation interval. This is particularly striking in dairy cattle. Until recently, selection of AI sires required the tedious, although highly accurate, progeny-testing procedure. In the latter, the breeding value of a candidate bull is evaluated from the milking performances of tens to hundreds of daughters. Obtaining a progeny-test for a given bull takes five to six years at an estimated cost of ~30,000 € per bull. In contrast, QTL information can be obtained at day one if not from a biopsy of a pre-implantation embryo.

Despite these theoretical benefits of MAS, its implementation has been relatively limited until recently despite the identification of hundreds of QTL affecting pretty much all traits of interest. This is primarily due to (i) the fact that most QTL have been mapped, at best fine-mapped, but that the QTN have remained elusive hence complicating the implementation of MAS, and (ii) the fact that the identified QTL only explain a small fraction of the genetic variation for the traits of interest.

The situation has dramatically changed, however, with the recent proposition of "Genomic Selection" (Meuwissen and Goddard, 2001). In this GS approach, "posterior" marker or haplotype effects are estimated in a "training" population composed of animals with both marker and phenotype data. This is achieved by simultaneously exploiting linkage (QTL mapping) and linkage disequilibrium (QTL fine-mapping) without consideration of statistical significance (yet shrinking the QTL effects towards prior values in the absence of overwhelming experimental support). All estimated effects are then used to predict the genomic breeding value of a tested individual.

To be effective GS requires high density marker maps covering the entire genome. This dream has recently become reality for a growing list of livestock species thanks to the availability of high-throughput sequencing technologies allowing for the rapid detection of millions of SNPs, as well as the development of high-throughput platforms allowing cost-effective genotyping of hundreds of thousands of SNPs in parallel.

GS represents a genuine revolution in animal breeding and is already being implemented by the most advanced AI companies in the world.

1.4. A paternally imprinted QTL with major effect on muscle mass and fat deposition maps in the vicinity of the porcine *IGF2* locus:

Prior to the initiation of this thesis, the previously described QTL mapping and fine-mapping principles were applied to a Piétrain x Large White intercross, resulting in the identification of a QTL with major effect on carcass composition on porcine chromosome 2 (Nezer et al., 1999 and 2002).

Piétrain pigs, originating from the village of Piétrain in Belgium, are well-known for their exceptional meatiness, unfortunately accompanied by modest growing performances as well as susceptibility to stress resulting in "pale-soft-exudative" meat as a result of a missense mutation (R615C) in the gene encoding the ryanodine receptor (Fujii et al., 1991). In many respect, Large White have complementary features, having fattier carcasses, being "stress-negative" and growing more rapidly.

QTL mapping conducted in a Piétrain x Large White population comprising 525 individuals phenotyped for 15 traits and genotyped for 137 microsatellite markers resulted in the identification of two significant QTL on, respectively, chromosomes 2 (muscle mass and fat deposition) and 7 (growth, carcass length and composition) and two suggestive QTL on, respectively, chromosomes 1 (fat deposition) and 13 (fat deposition) (Nezer et al., 2002).

In these experiments, the QTL on the distal end of the short arm of chromosome 2 (fig. 3) was by far the strongest, yielding lod score values > 20 and causing major effects on muscularity and fat deposition. Comparative mapping indicated that the corresponding chromosome region in human was HSA11p15.5, known to harbor the *IGF2* and *MYOD* genes. These were considered to be interesting candidates on the basis of their known role in myogenesis. Polymorphisms were identified in the porcine orthologues and genotyped in the F2 pedigree, to determine the position of the corresponding genes on the porcine linkage map. The position of *IGF2* proved to coincide exactly with the most likely position of the QTL.

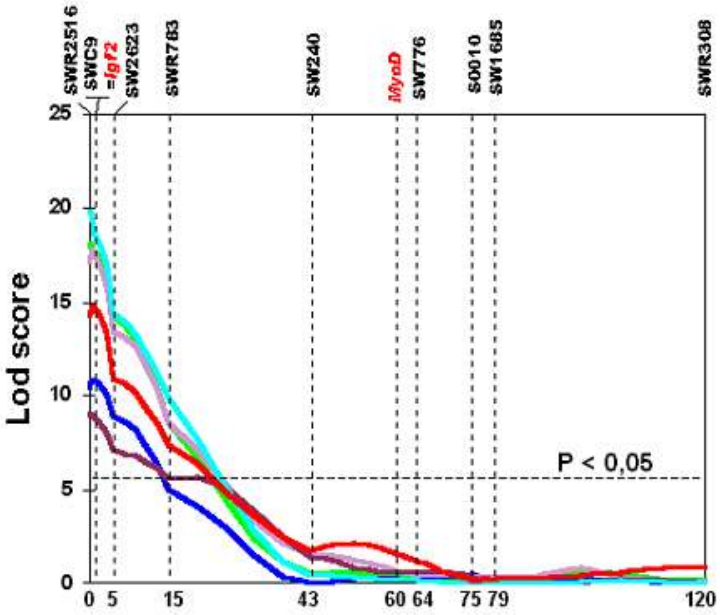


Figure 3: Lod score curves obtained in a Piétrain x Large White intercross for six phenotypes measuring muscle mass and fat deposition on pig chromosome 2. The most likely positions of the *IGF2* and *MYOD1* genes determined by linkage analysis, with respect to the microsatellite marker map, are shown (Nezer et al, 1999).

IGF2 was known to be imprinted in human and mice, with preferential expression from the padumnal allele (Giannoukakis et al., 1993; Tilghman et al., 1993). Assuming that the identified QTL was due to a QTN affecting *IGF2*, the QTL effect should likewise be preferentially associated with the segregation of the padumnal rather than madumnal homologue. This hypothesis was tested and spectacularly verified: lod scores increased when testing the effect of the padumnal homologues and maximized exactly at the *IGF2* position, while being zero when testing the madumnal chromosomes. The same results were confirmed

in an independent data set, resulting in the first description of an imprinted QTL (Nezer et al., 1999; Jeon et al., 1999).

IGF2 thus stood out as a perfect positional candidate gene. However, sequencing of the open reading frame of *IGF2* in the F1 generation did not reveal any underlying structural variant of *IGF2* that might explain the observations, nor even of any silent mutation.

Imprinted genes are known to occur in clusters encompassing multiple imprinted genes. Human chromosome 11 for instance harbors two such imprinted gene clusters including at least four transcripts preferentially expressed from the padumnal allele. Also, one could not exclude the existence of other, as of yet not identified imprinted genes in the vicinity. It is important to remember in this regard that, although the positions of *IGF2* and the QTL seemed to coincide, the confidence interval for the QTL spanned several tens of centimorgan. Moreover, even if *IGF2* were the causal gene, the QTN might affect regulatory elements of which some are known to be located at tens to hundreds of kilobases of the actual target gene.

It was thus decided to genetically fine-map the QTL before engaging in extensive sequencing and functional characterization of *IGF2*. This was achieved by following the MASA approach described above. A BAC contig spanning ~1 Mb encompassing the *IGF2* gene was constructed and 54 novel markers developed. Fourteen paternal half-sib pedigrees counting 819 offspring were phenotyped and genotyped for 31 chromosome 2 markers, resulting in the identification of five putative Q and six putative q chromosomes. These were genotyped for the newly developed high density map, resulting in the identification of a 250 Kb haplotype shared identical by descent by all Q chromosomes, hence positioning the QTN in this interval (fig. 4). Interestingly, the interval was bounded on the proximal side by SWC9 located in the 3'UTR of the *IGF2* gene, thus placing the QTN upstream of this recombinational breakpoint and hence excluding involvement of all regulatory elements located on the *H19* side of the candidate gene. Note that the interval encompasses *INS*, showing at least partial monoallelic expression in thymus (Vafiadis et al., 1997; Pugliese et al., 1997) and yolk sac (Gudrun et al., 2001) in human. A possible role for INS in mediating the QTL effect could thus not be excluded.

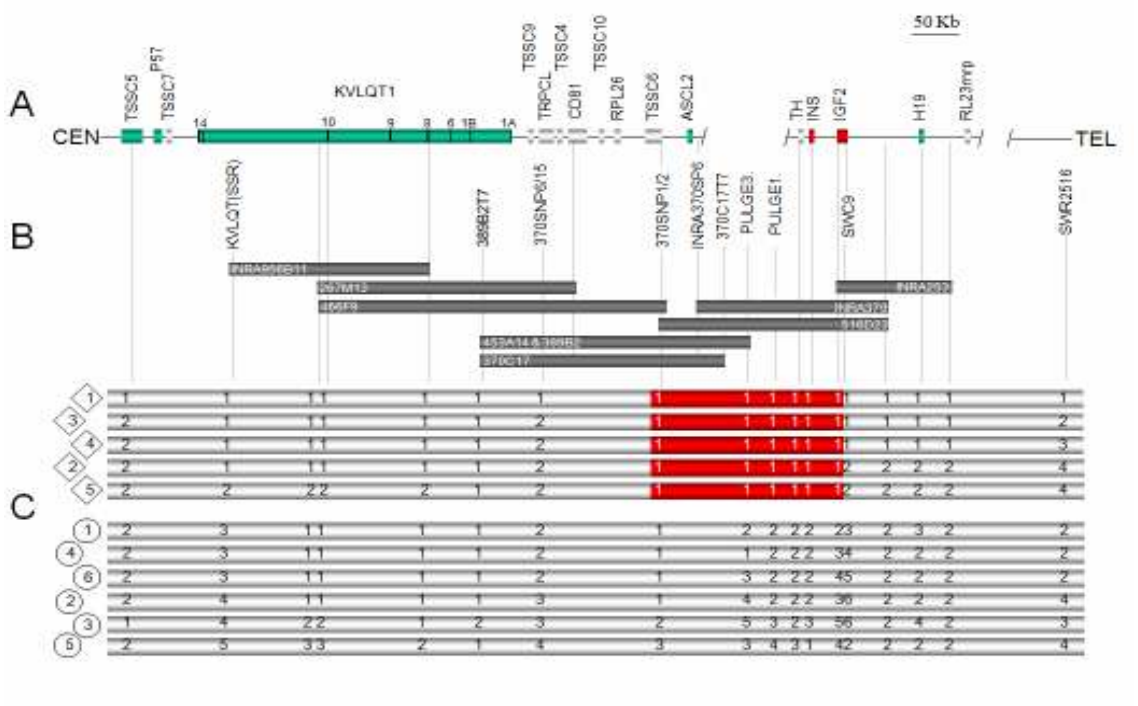


Figure 4: (A) Schematic representation of the human 11p15 imprinted domain according to Onyango *et al*. (2000). Biallelically expressed genes are shown as grey shaded cylinders. Maternally expressed genes are shown as green shaded cylinders and paternally expressed genes are shown as red shaded cylinders. (B) BAC contig spanning the porcine ortholog of the 11p15 imprinted domain, assembled by STS content mapping. The length of the horizontal bars does not reflect the actual physical size of the corresponding BACs. It is assumed in this map that the gene order with respect to telomere is conserved in man and pig (http://www.ensembl.org/homo_sapiens/). (C) Marker haplotypes of the five *Q* chromosomes (diamonds) and six *q* chromosomes (circles). Closely linked SNPs (<5 kb) or adjacent SNPs that could not be ordered were merged into polyallelic multisite haplotypes. The reddened chromosome segments correspond to the haplotype shared by all *Q* chromosomes and are therefore assumed to contain the QTL (Nezer et al., 2003).

1.5. Insulin-like growth factor 2 (IGF2):

1.5.1. Insulin-like growth factors:

Protein structure: As their name implies, the insulin like growth factors (IGF1 and IGF2) are structurally related peptidic hormones (~65% amino-acid similarity with each other) that resemble insulin in their primary sequence (~50% amino-acid similarity) and three-dimensional conformation (e.g. LeRoith & Bondy, 1996). The mature IGF1 and IGF2 peptides correspond to central peptides of ~65 residues devoid of the amino-terminal signal peptide (~25 a.a.) and carboxyterminal E-peptide (~90 a.a.) severed by proteolytic processing (posttranslational modification at Golgi apparatus). The mature IGFs adopt a conformation that is very similar to proinsulin. The C-peptide equivalent is shortened but complemented by an amino-terminal D-extension of the A-chain.

Pre- and post-natal growth promoting effects: The insulin-like growth factors (IGF1 and IGF2) regulate growth and development of multiple tissues during embryonic and foetal stage. IGF2 knock-out mice are characterized by proportionate dwarfism apparent from midgestation (i.e. after placentation) but essentially normal otherwise (deChiara et al., 1990; Baker et al., 1993), whereas increased levels of IGF2 expression in transgenic mice have been shown to cause fetal overgrowth (Eggenschwiler et al. 1997). IGF1 knock-out mice exhibit growth retardation in late gestation (without reduction in placental size) and most of them die shortly after birth (Liu et al., 1993; Powell-Braxton et al., 1994). IGFs are also known as "somatomedins" as they are secreted by various tissues in response to GH stimulation, mediating its growth-promoting effect. However, several tissues may secrete IGFs in the absence of GH stimulation. It is noteworthy that targeted deletion of the *IGF1* gene in liver demonstrated that liver-derived IGF1 is not required for postnatal body growth (Sjogren et al., 1999).

Endocrine, paracrine and autocrine mode of action: Circulating IGF levels are thought to primarily reflect synthesis by the liver. Circulating IGFs probably act on several target tissues in an endocrine way. In addition, many tissues (including skeletal muscle) secrete IGFs acting locally in a paracrine/autocrine fashion. IGFs have been dubbed "extracellular second messengers" mediating the action of many mitogenic stimuli (e.g. Florini et al., 1996).

Signaling and non-signaling receptors: Both IGF1 and IGF2 mainly signal via the IGF1 receptor. This widely expressed ligand-stimulated transmembrane tyrosine-protein kinase activates several intracellular signaling cascades through a series of adaptor molecules including insulin substrate-1 (IRS-1) (Wilson and Rotwein, 2006). In addition, IGF2 binds to a "type II" IGF2 receptor, corresponding to the cation-independent mannose-6-phosphate receptor. This IGF2/M-6-P receptor is non-signaling and acts as an antagonist targeting IGF2 for degradation in lysosomes.

IGF binding proteins: The majority of IGF molecules are not free but rather bound to one of six IGF binding proteins (IGFBP-1 to 6). IGFBP are secreted by many tissues including liver and regulate the biological action of IGFs by modulating bioavailability.

Effect on skeletal muscle: IGFs have been shown to sequentially promote multiplication and differentiation of several myoblasts lines in culture (e.g. Florini et al., 1996). IGF2 may capacitate the effect of the basic helix-loop-helix transcription factors of the MyoD family by targeting transcriptional co-regulators including p300 and P/CAF (Wilson & Rotwein, 2006). Mortality of *IGF1* and *IGF1-R* knock-out mice is thought to result from hypoplasia of

respiratory muscles. Transgenic overexpression of IGF1 in skeletal muscle causes pronounced muscular hypertrophy (Musaro et al., 2001).

1.5.2. The IGF2 gene:

Multiple IGF2 isoforms: The *IGF2* gene spans ~30 Kb on chromosomes 11p15 and 7 in human and mice, respectively. It comprises nine exons of which only the last three are protein-encoding. Its expression is controlled by at least five promoters (P0 to P4) that generate tissue-specific isoforms with distinct 5'UTR sequences. In addition, IGF2 antisense transcripts of unknown function have been described, as well as fusion transcripts with the adjacent *INS* gene (e.g. Monk et al., 2006).

Parental imprinting: Therians (comprising Marsupialia and Placentalia) forgo diploidy of ~100 of their genes by preferentially expressing only one of the two available alleles in a parent-of-origin specific fashion. For approximately halve of these "imprinted" genes the expressed allele is paternal, for the other halve maternal. Paternal and maternal alleles of imprinted genes carry distinct DNA methylation marks or imprints at Differentially Methylated Regions (DMRs). Part of those imprints are primary, established in the germline on so-called imprinting control regions (ICRs). For the majority of imprinted genes, primary methylation marks are established in the female germline, and will thus mark the maternal allele of the offspring. Primary imprints inherited via the gametes (whether oocyte or spermatozoa) resist the genome-wide demethylation that characterizes the first cleavage divisions of the embryo. Subsequently, secondary methylation marks are acquired somatically. In the soma, primary and secondary methylation marks are maintained throughout mitotic divisions by the DNA methyltransferase DNMT1. In the germline, however, imprints are first erased before reestablishment of novel sex-specific imprints by DNMT3A and its cofactor DNMT3L. Primary and secondary methylation marks affect the expression of imprinted genes either by modulating the function of insulators (see hereafter) or of promoters driving the expression of long non coding cis-silencing RNAs (e.g. Reik, 2001). Parental imprinting is thought to have emerged as a result of the unequal parental provision of resources to offspring and polygamy of most mammals. Under such circumstances, genes promoting the transfer of resources from mother to offspring benefit from mutations that silence the madumnal yet activate the padumnal allele, while the opposite applies to genes that oppose the action of such genes. This "parental conflict hypothesis" was to a large extend inspired by the opposite imprinting of the placental IGF2 growth factor (padumnal expression) and its IGF2/M-6-P receptor antagonist (madumnal expression).

The IGF2-H19 imprinted domain: The majority of imprinted genes occur in clusters, encompassing both paternally and maternally expressed genes. One of the best known imprinted domains is the *IGF2-H19* domain spanning about 150 Kb on chromosome 11p15 in human and on chromosome 7 in the mouse. *H19* encodes a highly conserved, non-coding pre-miRNA of unknown function (Smits et al., 2008). *IGF2* and *H19* share a series enhancers located distally from *H19*. Reciprocal imprinting (paternal expression of *IGF2* and maternal expression of *H19*) is thought to reflect an epigenetic switch that (i) on the *madumnal* allele, sequesters the *IGF2* promoters (except the biallelically expressed, liver-specific P1 promoter) in an inactive chromatin loop anchored in a CTCF-mediated interaction between the unmethylated primary DMR (located 2 Kb upstream of *H19*) and the secondary DMR1 located in intron 3 of *IGF2*, and (ii) on the *padumnal* allele, precludes CTCF-mediated formation of the inactive loop by methylation of DMR and DMR1 thereby allowing access of the *IGF2* promoters to the shared enhancers while silencing H19 by promoter methylation (Reik et al., 2004). Besides the common mechanism (interaction between H19 DMR, enhancers, DMR1, and IGF2 promoters) that regulates the reciprocal imprinting of IGF2 and H19, differential methylation of the DMR0, DMR1, and DMR2 may regulate IGF2 expression by distinct manners and independent of H19 DMR. In the mouse, DMR0 spans IGF2 promoter P0, exon 1 and intron 1. DMR1 and DMR2 cored in the intron 2 and 8, respectively. Maternal methylation of DMR0 may be sufficient to repress IGF2 expression from maternal allele. In contrast to DMR0, DMR1 and DMR2 are paternally methylated. The silence of IGF2 maternal allele was caused by the binding of GCF2, a silencer element, to the hypomethylated maternal allele (Constancia et al. 2000; Eden et al. 2001)) while methylation of paternal DMR2 increases IGF2 transcription (Murrell et al., 2001).

1.5.3. Phenotypic effects of germline and somatic mutations and epimutations in the IGF2-H19 domain:

Beckwith-Wiedemann syndrome (BWS)(e.g. Weksberg et al., 2003; Rahman, 2005; Delaval et al., 2006): Gross alterations in IGF2 dosage, resulting from germline or somatic mutations and epimutations, causes severe pathological conditions. Generalized overexpression of IGF2 underlies a subset of cases of BWS, characterized by pre- and postnatal overgrowth, macroglossia, abdominal wall defects, organomegaly, hemihyperplasia, neonatal hypoglycemia, ear abnormalities and increased risk of Wilm's tumor of the kidney. Most BWS cases are sporadic, resulting either (i) from paternal uniparental disomies (pUPD) encompassing the *IGF2-H19* and neighboring *KCNQ1* imprinted domains, (ii) from

chromosomal aberrations perturbing imprinted expression in the region, (iii) from somatic loss-of-function mutations in the madumnal *CDKN1C* allele, (iv) from somatic epimutations affecting the ICR of either the *IGF2-H19* (acquisition of a padumnal epigenotype by the madumnal allele (by methylation gain) resulting in biallelic *IGF2* expression and extinction of *H19*) or *KCNQ1* domain (acquisition of a padumnal epigenotype by the madumnal allele (by methylation loss) causing biallelic expression of the *KCNQ1OT1* non-coding cis-silencing transcript and extinction of the *CDKN1C* (=$p57^{KIP2}$) growth inhibitor), or (v) from unidentified molecular causes accompanied or not by loss of imprinting in either domain. In rare cases, BWS is familial with parent-of-origin dependent autosomal dominant inheritance. Such familial cases have been shown to involve loss-of-function mutations of *CDKN1C* (e.g. Hatada et al., 1996; Niemitz et al., 2004), as well as microdeletions of the *IGF2-H19* ICR causing conferring a paternal epigenotype on maternal transmission (e.g. Sparago et al., 2004). Increased risk for Wilm's tumor is restricted to BWS cases with demonstrated overexpression of *IGF2* (Rahman, 2005).

Silver-Russell syndrome (SRS) (e.g. Eggerman et al., 2008): SRS is characterized by intrauterine and postnatal growth retardation, facial dysmorphism and a series of minor features including relative macrocephaly, skeletal asymmetry, precocious puberty and genital abnormalities. SRS was initially associated with maternal UPDs of chromosome 7 encompassing the imprinted, maternally expressed *GRB10* growth repressor. More recently, the occurrence of genetic and epigenetic alterations in the *KCNQ1* and *IGF2-H19* imprinted domains on chromosome 11 were demonstrated in a sizeable (~50%) proportion of SRS cases as well. Hypomethylation of the *IGF2-H19* ICR in particular, causing down-regulation of IGF2 expression, was most commonly observed.

Childhood and colorectal cancers (e.g. Feinberg & Tycko, 2004): An oncogenic role for imprinted genes mapping to 11p15 was provided by the discovery of the systematic retention of the paternal allele in Wilms' tumors and embryonal rhabdomyosarcoma with loss-of-heterozygosity (LOH) at 11p15. While the most obvious explanation is the loss of anti-oncogene expressed exclusively from the madumnal allele, an alternative explanation is that several tumors were actually isodisomic (as demonstrated for rhabdomyosarcoma) leading to overexpression of a padumnally expressed "proto-oncogene", possible *IGF2*. Supporting this hypothesis is the subsequent discovery of loss-of-imprinting (LOI) of the *IGF2* gene in a significant proportion of Wilms' tumors, often accompanied by hypermethylation of the *IGF2-H19* ICR. Interestingly, *IGF2* LOI was subsequently observed in colonic mucosa of ~10% of healthy people, yet to be a major risk factor for the subsequent development of

colorectal cancer (Cui et al., 2003). Moreover, aberrant methylation of the IGF2-H19 ICR was shown to be characterized by familial clustering supporting inherited tendency for LOI (Sandovici et al., 2003).

Common complex diseases: Common variants in the *IGF2* gene have been claimed to be associated with body weight, body mass index and height, but (as for many candidate gene studies) these findings have been difficult to replicate (e.g. Heude et al., 2007) and have not been confirmed by genome wide association studies (GWAS). Interestingly, padumnal class I allele at the INS VNTR has been associated with childhood obesity (Le Stunff et al., 2001). Genetic variation in this VNTR is known to affect expression levels of insulin but also *IGF2*.

CHAPTER 2

COMPARATIVE SEQUENCE ANALYSIS OF THE *INSULIN-IGF2-H19* GENE CLUSTER IN PIGS

Valérie Amarger[1,3*], Minh Nguyen[2*], Anne-Sophie Van Laere[1], Martin Braunschweig[1], Carine Nezer[2], Michel Georges,[2] and Leif Andersson[1]

Mammalian Genome, Volume 13(7):388-398, 2002

[1]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 597, SE-751 24 Uppsala, Sweden;

[2]Department of Genetics, Faculty of Veterinary Medicine, University of Liège (B43), 20, bd. de Colonster, 4000 Liège, Belgium;

[3]Present address: UMR1061 INRA/Université Limoges, 123 av. Albert Thomas, 87060 Limoges, France.

*These two authors contributed equally to this work

Minh Nguyen's contribution:

1. Construction of the Not1 restriction map of two pig BAC clones
2. DNA sequencing and sequence analyses of the *H19* region
3. Contribution to DNA sequencing of the *IGF2* region
4. Contribution to sequence annotation

Summary

IGF2 is the major candidate gene for a paternally expressed Quantitative Trait Locus (QTL) in the pig primarily affecting muscle development. Here we report two sequence contigs together comprising almost 90 kb containing the *INS – IGF2* and *H19* genes. A comparative sequence analysis of the pig, human, and mouse genomic sequences was conducted to identify the exon/intron organization, all promoters, and other evolutionarily conserved elements. RT-PCR analysis showed that *IGF2* transcripts originated from four different promoters and included various combinations of seven untranslated exons together with the three coding exons, in agreement with previous findings in other mammals. The observed sequence similarity in intronic and intragenic regions among the three species is remarkable and is most likely explained by the complicated regulation of imprinting and expression of these genes. The general trend was, as expected, a higher sequence similarity between human and pig than between these species and the mouse, but a few exceptions to this rule were noted. This genomic region exhibits several striking features, including a very high GC content, many CpG islands, and a low amount of interspersed repeats. The high GC and CpG content were more pronounced in the pig than in the two other species. The results will facilitate the further characterization of this important QTL in the pig.

2.1. Introduction:

A paternally expressed QTL (Quantitative Trait Locus) affecting muscle mass in pig has been identified at the distal end of pig Chromosome (Chr) 2p. The QTL was mapped independently in a Large White/Pietrain intercross (Nezer et al. 1999), a Wild Boar/Large White intercross (Jeon et al. 1999), and later in an intercross between Landrace/Large White and Meishan pigs (de Koning et al. 2000). Pig Chr 2p1.7 shows conserved synteny with human Chr 11p15, which is extensively studied because of the presence of a cluster of imprinted genes. Among them, the insulin-like growth factor II (*IGF2*) gene is paternally expressed and it was identified as the major candidate gene for the QTL, because of its involvement in muscle growth and differentiation (Florini et al. 1995). The paternal expression of pig *IGF2* has been confirmed (Nezer et al. 1999).

IGF2 is flanked on its 5' and 3' sides by the insulin *(INS*) and *H19* genes, respectively, and these three genes cover a region of about 150 kb on human Chr 11p15 and mouse Chr 7 (Zemel et al. 1992; Onyango et al. 2000). These three genes seem to have a closely related regulation and have been extensively studied because of their involvement in several pathologies. The VNTR present in the 5' region of the human *INS* gene is associated with

susceptibility to insulin-dependent diabetes mellitus (IDDM; Bennet et al. 1995). This VNTR has an effect on *INS* mRNA levels (Pugliese et al. 1997; Vafiadis et al. 1998) and it also influences the expression of *IGF2* in human placenta *in vivo* (Paquette et al. 1998). However, this transcriptional effect is absent in leukocytes (Vafiadis et al. 1998), suggesting a tissue-specific regulation dependent on the particular promoter used for *IGF2* transcription. *IGF2* is a complex transcription unit that consists of 10 exons in human (Mineo et al. 2000). The first seven exons (denoted 1-6 and 4b) are non-coding leader exons while exons 7-9 encode pre-pro IGF2 consisting of 180 amino acid residues. Exons 1, 4, 5, and 6 are preceded by distinct promoters (P1-P4) which give rise to a family of mRNA transcripts containing different leader exons but the same coding exons (Holthuizen et al. 1990). The different promoters confer a tissue-specific as well as a developmental-specific expression of the gene.

IGF2 and *H19* are expressed in a monoallelic fashion from the paternal and maternal chromosomes, respectively, and their imprinting is closely co-regulated. Over-expression of *IGF2*, with or without disruption of the imprinting pattern of itself and *H19* is implicated in several disorders in the human, mostly growth disorders and tumors. For instance, the Beckwith-Wiedemann syndrome shows evidence of both *H19*-dependent and *H19*-independent pathways affecting the *IGF2* imprinting status (Brown et al. 1996, Reik et al. 1995).

Considering the complex regulation of *IGF2* and the close interaction between *INS*, *IGF2*, and *H19*, our first step in understanding the molecular basis of the QTL effect was to sequence the region covering the three genes in pig. Because no difference in the coding sequence of *IGF2* was identified in animals carrying different QTL alleles, the causative mutation(s) may be regulatory (Nezer et al. 1999). Comparative sequencing is a powerful tool to identify functionally important sequences that are evolutionarily conserved even between distantly related organisms. Human and mouse comparative sequence analysis was used to identify new genes and potential regulatory elements in the human Chr 11 imprinted domain (Onyango et al. 2000; Ishihara et al. 2000). In this study, we used comparative sequence analysis of pig, human, and mouse to define the organization of these three genes in pig and to identify potential regulatory elements that could be responsible for the QTL effect.

2.2. Materials and methods:

2.2.1. Human and mouse sequence data:

The human sequence covering the *INS-IGF2* regions is a combination of two overlapping sequences available in Genbank: L15440 (from 1 to 12348) and AC006408 (from 69001 to 42189, reverse complemented). The mouse sequence covering this region was taken from the sequence AC012382. The human *H19* sequence is from AC004556 and the mouse *H19* sequence from AP003182 and AF049091.

2.2.2. Restriction mapping of the BAC clones:

BAC DNA was purified using the QIAGEN plasmid midi kit (QIAGEN, Germany). Two µg of BAC DNA were digested with 10 units *Not*I restriction enzyme. The fragments were separated by PFGE. Gels were run at 4 V/min for 16 h at 14°C with the pulse times ramped from 0.1 to 2.5 sec. Following electrophoresis, gels were stained with ethidium bromide, and the fragments were visualized by exposure to UV light. *Not*I restriction fragments were subcloned into pNEB193 (New England Biolabs).

2.2.3. BAC sequencing:

DNA from BAC 370 was purified by an alkaline lysis method followed by phenol/chloroform extraction. Twenty mg of DNA was partially digested by 10 units *Sau*3AI (New England Biolabs) for 10 min at 37°C. Digested DNA was separated on 1% agarose gels, fragments between 1.5 and 2.5 kb were excised and purified using QIAEX II kit (QIAGEN, Germany). About 100 ng of purified DNA was ligated into 100 ng of *Bam*H1 restricted pUC18 (Amersham-Pharmacia Biotech, Uppsala, Sweden) by using T4 DNA ligase (New England Biolabs) and was used to transform XL1 blue *E. coli*. Plasmid DNA was prepared by an alkaline lysis method and sequenced with universal M13 reverse and forward primers by using the Big Dye Terminator sequencing kit (Perkin Elmer Applied Biosystem). Sequences were run on an ABI 377 automatic sequencer (Perkin Elmer Applied Biosystem). Difficult templates with a very high GC content and long stretches of Gs or Cs were sequenced by using the dGTP Big Dye Terminator kit (Perkin Elmer Applied Biosystem).

*Not*I restriction fragments containing the *H19* region were sequenced using the EZ::TN™ Transposon Insertion System (Epicentre Technologies, Madison, WI). Transposon inserted recombinant plasmid DNA was sequenced as described above.

2.2.4. Bioinformatic analysis:

Sequences were assembled using the Phred/Phrap/Consed package (Ewing et al. 1998; Gordon et al 1998). The assembled sequences were then analyzed with a variety of computer software programs. Sequence comparison with cDNA sequences was done with pairwise BLAST at http://www.ncbi.nlm.nih.gov. Repetitive elements were localized and identified by RepeatMasker (A.F.A. Smit and P. Green, unpublished; http://ftp.genome.washington.edu/index.html). In order to detect pig specific interspersed repeats, the mammalian library of repeats provided with the program was updated with a consensus pig SINE sequence and other pig specific repeats. Sequence identity plots were obtained by using VISTA (Dubchak et al. 2000; Mayor et al. 2000) at http://www-gsd.lbl.gov. The comparison between pig and human sequences was done with Alfresco (Jareborg and Durbin 2000). Alfresco uses the program CpG (G. Micklem and R. Durbin, unpublished) for determining the presence of CpG islands. By default, a CpG island is defined as a DNA stretch at least 200 bp long with a GC content >50% and an observed to expected ratio of CpG dinucleotides > 0.6 (Gardiner-Garden and Frommer 1987). Conserved elements were identified by using DBA (included in Alfresco) and pairwise BLAST as said above.

2.2.5. RT-PCR analysis of IGF2 transcripts:

Adult and fetal tissue samples were immediately frozen in liquid nitrogen and stored at –70^{o}C or in RNAlater™ (Ambion) until total RNA was prepared by using TRIzol (GIBCO BRL) according to the manufacturer's protocol. The isolated RNA was DNase I (Ambion) treated, which was subsequently inactivated by phenol-chloroform extraction. First-strand cDNA synthesis was done using total RNA samples following the manufacturer's instructions (Amersham Pharmacia Biotech).

RT-PCR was carried out with the Advantage®-GC cDNA PCR kit (CLONTECH). The following primers were used to determine the usage of the four promoters (P1-P4): P1, forward primer *IGF2*EX1F 5'-GGTAGGCGGCTGGGATGAGTGG-3' in exon 1 and reverse primer *IGF2*EX8R 5'-TGCCGGCCTGCTGAAGTAGAAG-3' in the junction between exon 7 and 8; P2, forward primer *IGF2*EX4F 5'-TCCCTGGGTCTTCCAACG-GACTGGGCGT-3' in exon 4 and reverse primer *IGF2*EX7R 5'-CTCACTGGGGCGGTAAGCAGCATAGCAG-3' in exon 7; P3, forward primer *IGF2*EX5F 5'-CGGCCCGTCCTCCCCAAACAATCAGA-C-3' in exon 5 and reverse primer *IGF2*EX7R 5'-GGGCG-GTAAGCAGCATAGCAGCA-CGAG-3' in exon 7; and forward primer *IGF2*EX6F 5'-GGCAGGCTCCCAGCTTCCTCCT-CCTCC-3' in exon 6 and the reverse primer *IGF2*EX9R 5'-GGGCGGACTGCTT-

CCAGGTGTCATAGC-3' in exon 9 for P4. The obtained PCR products were isolated from agarose gels and sequenced directly on a MegaBACE™ 1000 sequencing instrumentally, using the DYEnamic™ ET dye terminator cycle sequencing kit (Amersham Pharmacia Biotech).

2.3. Results:

2.3.1. Restriction mapping of the pig BAC clones:

Two BAC clones (BAC 253 and 370) containing *IGF2* were isolated from a pig genomic library using *IGF2* primers (Jeon et al. 1999). DNA from these two clones was digested with the restriction enzyme *Not*I and the resulting restriction fragments separated by conventional as well as pulsed field gel electrophoresis (PFGE). BAC 253 and 370 contained nine and seven *Not*I restriction fragments, respectively, ranging in size from 1.2 to 63 kb. All these fragments (except the 63 kb fragment) were subcloned and sequenced from both ends. Outward pointing primers were designed for all subclones and used for sequencing with the BAC DNA as template. Comparison of the obtained sequences and the end sequences of the subclones allowed unambiguous ordering of all *Not*I restriction fragments (Fig. 1A). The location of the *INS* and *H19* genes in BAC 370 and 253 was established by PCR amplification and sequencing. The tyrosine hydroxylase (*TH*) gene was identified during the sequencing process.

2.3.2. Sequencing of the INS-IGF2 and H19 regions:

A shotgun library was constructed for BAC 370 containing the *INS-IGF2* region. One thousand clones were sequenced from both ends, giving approximately 1,600 high-quality sequences. After assembly, three strategies were used to fill gaps. The first strategy involved a simple primer walk when a gap was found between the two ends of a shotgun clone. The second strategy involved PCR amplification (using the BAC DNA as template) of a gap situated between two contigs that could be ordered and oriented after a comparative analysis using the homologous human sequence. The third strategy involved subcloning of BAC restriction fragments covering a gap followed by sequencing by primer walk. The 24 kb *Not*I fragment containing *H19* as well as the 1.2, 1.5, 6.1, and 24 kb *Not*I fragments containing the *H19* upstream region were subcloned and sequenced using transposon insertions. A 32 kb contiguous sequence containing the last five exons of *TH* and the complete *INS* and *IGF2* genes and a 56 kb sequence containing the *H19* gene were determined. All the *Not*I sites of

the restriction map were found in the sequence, allowing a precise localization of the genes on the restriction map (Fig. 1). The distances between the genes were determined as follows: *TH* - 1.9 kb - *INS* - 0.7 kb - *IGF2* - 88.1 kb - *H19*.

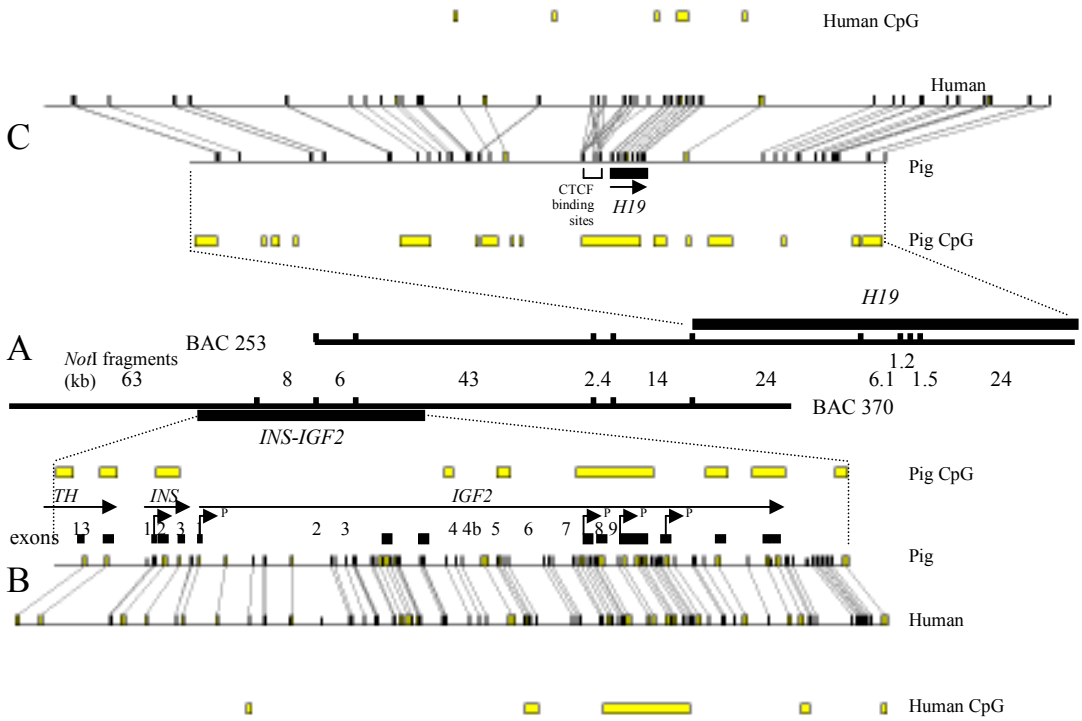


Fig. 1: A: NotI restriction map of two pig BAC clones (253 and 370) including the *INS-IGF2* and *H19* regions. B and C: Comparative sequence analysis of the pig and human sequences using Alfresco. Conserved elements are represented by two boxes linked by a line. Shaded and solid boxes are conserved elements identified by BLAST and DBA algorithms, respectively. Positions of exons, promoters, and CpG islands are indicated.

The GC content of the two sequenced regions is significantly higher in pig than in the corresponding region in human (Table 1) and in both species much higher than the genome average. A large number of CpG islands was also identified and, in line with the difference in the GC content, the number and sizes are larger in pig than in human. This is most pronounced in the *H19* region, where the total length of CpG islands is about 10 times higher in pig than in human (Table 1; Fig. 1).

The sequences were screened for repetitive sequences by using RepeatMasker (Table 1). The total number of repeats is higher in human than in pigs. In the *H19* region, this difference is due only to the proportion of interspersed repeats, which is more than four times higher in human than in pig, whereas the number of simple repeats is similar. The *INS-IGF2* region is characterized by a surprisingly low amount of interspersed repeats and a high proportion of

simple repeats in both species. The only interspersed repeat found in the pig *IGF2* gene was a Mammalian Interspersed Repeat (MIR), and its location was conserved between pig and human; MIR elements represent a class of short interspersed repeats (SINE) found in all mammals.

Table 1: Global sequence comparison of the human and pig *INS-IGF2* and *H19* regions

	INS –IGF2		*H19*	
	Pig	Human[a]	Pig	Human[b]
Length (bp)	32,467	35,647	56,404	70,750
% GC	65.0	61.4	67.5	61.6
CpG islands	9	5	16	6
Total length of CpG islands (bp)	8,605	4,827	17,622	1,940
Total number of CpG dinucleotides	823	509	1,246	175
Interspersed repeats (% of total sequence)	0.47	1.41	1.66	8.21
Simple repeats (% of total sequence)	3.97	4.51	1.29	1.24
Total repeats (% of total sequence)	4.44	5.92	2.95	9.45

[a] The human *INS-IGF2* sequence is a combination of two overlapping sequences: L15440 (1 to 12348) and AC006408 (45702 to 69001, reverse complementary strand)

[b] Genbank sequence AC04556 from 25967 to 96716, reverse complementary strand.

2.3.3. Structure of the pig INS, IGF2, and H19 genes:

The exon/intron organization of the pig genes was deduced by aligning the genomic sequence with cDNA sequences from pig when available (*INS* mRNA AF064555; *IGF2* mRNA X56094 and RT-PCR products described below) or from human (*H19*). We identified 10 *IGF2* exons in the pig and the corresponding ten exons of human *IGF2* were identified by a combination of several mRNA sequences obtained from different tissues (GenBank M22372, X06259, X03423, Y13633, X56539, X56540, and X03562). The *H19* exons in pig were identified by sequence similarity with the human mRNA sequence (GenBank M32053). Exon and intron sizes, and sequence identities between the human and pig genes are presented in Table 2. *IGF2* exons 7, 8, and 9 are coding and evolutionarily well conserved. However, the seven untranslated exons are also very well conserved, with sequence identities ranging from 74 to 91%. The *insulin* gene, although all exons are coding, is less conserved than *IGF2,* and the first exon, encoding the signal peptide, is only 61% identical between the two species. The structure of *H19* is conserved but the level of sequence identity is on average lower than it is for *IGF2*. The level of sequence identity of intronic sequences shows a considerable variation.

Some introns are as conserved as exons, such as, for example, *INS* intron 1 (70%), *IGF2* introns 2, 3, 4, and 5 (68, 66, 74, and 73%, respectively).

Gene Exon/Intron	Exon length (bp) Pig	Human	Pairwise align. score[1] (%)	Intron length (bp) Pig	Human	Pairwise align. score[1] (%)
INS						
1	43	42	61	162	179	70
2	203	204	81	391	787	<50
3	140	146	85			
IGF2						
1	112	117	83	7410	8922	<50
2	193	220	83	1260	1318	68
3	232	242	74	6157	6319	66
4	389	390	78	536	557	74
4b	165	165	90	648	692	69
5	1148	1164	89	859	941	73
6	115	100	90	1817	1658	<50
7	163	163	88	1909	1705	<50
8	145	145	91	258	293	<50
9	240	237	88			
H19						
1	1354	1328	76	88	96	<50
2	126	135	77	78	95	51
3	118	113	54	83	80	<50
4	128	123	82	82	81	56
5	550	614	70			

Table 2: Comparative exon/intron organization of the human and pig *INS*, *IGF2*, and *H19* genes

[1]Pairwise alignment scores were determined by using ClustalW at http://www.ebi.ac.uk; ClustalW could not perform correct alignments for sequence identities <50%.

The human promoter sequences for *INS*, *IGF2*, and *H19* are available in the Eukaryotic Promoter Database (http://www.epd.isb-sib.ch/; Perier et al. 2000). We have used these sequences to identify the pig promoters by sequence similarity search. The positions of the promoters are conserved, and they all show a high sequence identity to the human ones (Fig. 2).

Insulin

```
human   GGGAGATGGGCTCTGAGACTATAAAGCCAGCGGGGGCCCAGCAGCCCTCagccctcc
pig     ..CGCCG...GG.A.GCG..........G..C...-...----------.......t
```

IGF2-P1

```
human   CCCGCCTCCAGAGTGGGGGCCAAGGCTGGGCAGGCGGGTGGACGGCCGGacactgga
pig     ......C.TG...A..A.....G..................C..............c
```

IGF2-P2

```
human   --AGAACTCTGCCTTGCGTTCCCCAAAATTTGGGCATTGTTCCCGGCTCGCcggccacc
pig     AGGCTG.....A.....A...--AGGCC................C..C.....cggt.t
```

IGF2-P3

```
human   CCTGGGCCGCGGGCTGGCGCGACTATAAGAGCCGGGCGTGGGCGCCCGCAgttcgcct
pig     .GG.C.....A..G.......G.............-...T...-....G.........
```

IGF2-P4

```
human   TGGGAGGAGTCGGCTCACACATAAAAGCTGAGGCACTGACCAGCCTGCAaactggac
pig     .........C.......G.....................................
```

H19

```
human   TTCTGGGCGGGGCCACCCCAGTTAGAAAAAGCCCGGGCTAGGACC-GAGGagcagggt
pig     ..................T.....................G..C.G.A........
        TTCTGGGCGGGGCCACCCCTGTTAGAAAAAGCCCGGGCTAGGGCCCGGGAagcagggt
```

Fig. 2: Sequence alignment of the human and pig promoter regions for the *INS*, *IGF2*, and *H19* genes. The lower case letters mark the transcription start. Human promoter sequences were found in the Eukaryotic Promoter Database (http://www.epd.isb-sib.ch). Identities to the human sequences are indicated by dots and alignment gaps are indicated by dashes.

2.3.4. Characterization of IGF2 transcripts and promoter usage in fetal and adult tissue:

The *IGF2* transcripts and the promoter usage for different adult and fetal porcine tissues documented by RT-PCR analysis are compiled in Table 3. Promoter 1 (P1) usage was predominantly detected in adult liver, fetal ham, and fetal liver. In addition to the transcript documented in the previously reported *IGF2* cDNA sequence (X560940) containing exons 1,3 and 7-9, we found a P1 transcript without exon 3 and a P1 transcript including exon 2. The P1 usage in fetal ham but not in adult muscle indicates a developmental specific usage of P1. The same applies to the two P1 transcripts found in adult liver, but not in fetal liver. Promoter 2 to 4 (P2-P4) transcripts were detected in all investigated tissues. P2 usage resulted in two different transcripts and the results showed that the transcript including exon 4b is much less abundant than the transcript without it. The sequencing results of the three pooled PCR products from adult muscle, liver, and kidney were consistent with the assumption that these transcripts containing exon 4b are identical. The finding of an *IGF2* exon 4b in pig agrees well with the results published by Ohlsen et al. (1994) for sheep and by Mineo et al. (2000), who confirmed the existence of a 10^{th} exon of *IGF2* in human.

Table 3: Characterization of porcine *IGF2* transcripts and promoter (P1-P4) usage in adult and fetal tissues by using RT-PCR[a]

	IGF2 promoter usage						
	P1			P2		P3	P4
Tissues	Exons 1,7±9	Exons 1,3,7±9	Exons 1±3,7±9	Exons 4,7±9	Exons 4,4b,7±9	Exons 5,7±9	Exons 6,7±9
Adult muscle		(+)		+	{+}	+	+
Adult liver	+	+	+	+	{+}	+	+
Adult kidney		(+)	(+)	+	{+}	+	+
Fetal ham	(+)		+	+	(+)	+	+
Fetal liver			+	+	(+)	+	+
Fetal kidney				+	(+)	+	+
Fetal heart				+	(+)	+	+
Fetal brain		(+)	(+)	+	(+)	+	+
Fetal placenta		(+)		+	(+)	+	+
Fetal lung				+	(+)	+	+

[a] The reverse PCR primers were located at the junction of exon 7 and 8 for the P1 amplicon, in exon 7 for the P2 and the P3 amplicons, and in exon 9 for the P4 amplicon. We assume that all transcripts contain the coding exons 7-9.

(+) = very faint PCR product of the corresponding size was obtained but not sequenced.

{+}= PCR products were pooled together and sequenced.

2.3.5. Comparative analysis of the putative Nctc1/Rhit1 and Ihit1 transcription units:

A mouse transcription unit mapping in the *Mrp123-H19* interval (~17 kb downstream of *H19*) was previously reported by Ishihara et al. (1998) on the basis of homologies with ESTs that were subsequently confirmed by Northern blot and RACE analyses. It was referred to as *Nctc1* by Ishihara et al. (2000) and subsequently as *Rhit1* by Onyango et al. (2000). The corresponding transcripts are non-imprinted, non-coding, and primarily expressed in skeletal muscle. The *Nctc1* transcription unit is characterized by an upstream exon of about 200 bp separated by an intron of 2235 bp from a downstream exon of at least 2474 bp. The 3' end of this exon, however, does not exhibit a consensus polyadenylation signal and is bounded in the mouse by a genomic poly-A track, which is predicted to have primed the reverse transcription step. It is, therefore, unlikely to correspond to the genuine 3' end of these transcripts.

The 5' end of *Nctc1* corresponds to an evolutionary footprint conserved between mouse, human, and pig (no. 93-97 in Table 4). However, with the exception of a human EST (AF313096) of unknown origin reported by Onyango et al. (2000), no other human or pig EST mapping to this region could be identified. The AF313096 EST spans the exon 1-intron 1

boundary and would, therefore, correspond to an unspliced message. It is noteworthy in this regard that the corresponding donor splice site is not conserved in the human. BLAST searches performed with the AF313096 EST revealed a very significant, perfect 144-bp match with exon 13 of a gene predicted in silico to be transcribed from the *Nctc1* antisense strand. This putative XM_073653 gene comprises 14 exons spanning about 20 kb between *H19* and *Mrp123*. Alignment of the corresponding exons with the corresponding human and pig sequences did not reveal any evidence for conservation of the corresponding intron-exon boundaries or open reading frame. We therefore suspect that XM_073653 is a false-positive gene prediction.

Onyango et al. (2000) also identified a 282-bp open reading frame about 21 kb upstream of *H19*. The corresponding murine DNA sequence was reported to reveal strong 1.0-kb transcripts in murine and human liver, placenta, and brain (human), by Northern blot analysis. It was referred to as *Ihit1*. The corresponding sequence is, however, conserved neither in the human nor in the pig. Genescan analysis of the corresponding regions in human, mouse, and pig did not provide evidence for statistically well-supported, evolutionarily conserved exons. *Ihit1* is, therefore, considered as a false positive gene prediction as well.

Table 4: Conserved elements (outside exons, promoters and simple repeats) found in the *INS-IGF2* (1 to 59) and *H19* (60 to 97) regions. Positions of *INS*, *IGF2,* and *H19* exons are indicated by arrows.

	Conserved elements	Position in pig sequence	Length bp	Sequence Identity %	Comment
INS – IGF2 region					
	1	3997-4014	18	100	mouse[3]
	2	4070-4116	47	89	mouse[3]
INS →	3	5536-5557	22	90	
IGF2 ex1 →	4	5574-5621	48	89	
	5	6945-7066	122	87	
	6	8066-8099	34	94	mouse[3]
	7	8531-8553	23	91	mouse[3]
	8	8600-8652	53	94	mouse[3]
	9	9660-9749	90	89	MIR repeat
	10	11303-11350	48	89	
	11	11449-11487	39	94	
	12	11869-11904	36	91	
	13	12185-12206	22	95	
	14	12220-12236	17	100	
	15	12279-12295	17	100	
	16	12955-12984	30	87	
	17	13038-13068	31	90	
ex2 →	18	13221-13379	159	85	mouse[3]
	19	13710-13771	62	83	
	20	13897-13976	80	90	
	21	14038-14059	22	91	Similar to mouse anti-sense transc. AB030734
ex3 →	22	14089-14131	43	88	
	23	15994-16023	30	90	
	24	16358-16398	41	95	
	25	16640-16667	28	89	
	26	17406-17715	308	84	mouse[3]
	27	18069-18092	24	91	CpG island (mouse DMR1)[3]
	28	18149-18263	115	84	
	29	18326-18345	20	95	
	30	18523-18553	31	97	
	31	19975-19997	23	100	
ex4 →	32	20196-21231	36	89	
	33	20346-20371	26	92	
	34	21171-21209	39	90	mouse[3]
	35	21884-21923	40	87	CpG island IGF2 intron 4
	36	22041-22093	53	100	
ex4b →	37	22119-22135	17	100	
ex5 →	38	24363-24429	67	85	*IGF2* intron 5
	39	24500-24540	41	93	
	40	24659-24695	37	89	
	41	24756-24793	38	95	
	42	24907-24938	32	92	
ex6 →	43	25339-25373	35	91	
	44	26059-26090	32	91	
	45	26162-26183	22	95	
	46	26536-26588	53	89	mouse[3]
ex7 →	47	28303-28367	65	84	
ex8,9 →	48	29880-29909	30	93	
	49	29947-29978	32	94	*IGF2* 3' UTR
	50	30135-30173	39	92	
	51	30945-30966	22	100	
	52	31123-31160	38	95	
	53	31203-31229	27	96	
	54	31333-31349	17	100	
	55	31377-31394	18	94	
	56	31508-31526	19	95	
	57	31584-31655	72	88	
	58	31736-31782	47	83	
	59	32180-32438	259	85	
H19 region					
	60	2099-2119	21	100	
	61	2258-2326	69	86	
	62	3997-4074	78	95	
	63	9333-9798	65	86	
	64	9858-9893	36	94	
	65	10847-10875	29	96	
	66	10933-11029	97	91	
	67	16095-16139	45	89	
	68	16215-16244	30	90	
	69	18473-18510	38	90	
	70	19276-19371	96	84	Similar AF313051[1]
	71	20266-20270	45	84	
	72	20899-21030	124	92	Similar AF313050[1]
	73	21232-21255	24	96	
	74	22416-22498	83	88	
	75	22585-22609	25	92	
	76	22701-22744	44	93	
	77	24280-24359	80	90	
	78	25472-25758	287	86	
H19 →	79	34222-34252	31	93	
	80	40176-40487	312	86	Enhancer I [3]
	81	42616-42771	156	81	Enhancer II[3]
	82	46454-46525	72	86	CS5[2,3]
	83	47476-47528	53	87	mouse[3]
	84	48223-48259	37	95	
	85	49281-49316	36	97	CS6[2,3]
	86	49351-49397	47	91	
	87	49495-49557	63	84	
	88	50758-50794	37	92	
	89	51368-51398	31	97	CS7 [2,3]
	90	51416-51475	60	89	
	91	52142-52196	55	95	CS8 [2,3]
	92	52194-52257	64	95	
	93	52445-52581	137	82	Similar AF313096[1]
	94	52617-52635	19	100	
	95	54991-55021	31	87	mouse[3]
	96	55070-55115	46	87	
	97	56348-56386	39	87	

[1]human / mouse conserved elements (Onyango et al. 2000), [2]conserved putative enhancers (Ishihara et al. 2000),
[3]human / mouse conserved elements (this study).

2.3.6. Highly conserved non-coding elements:

Global sequence identity plots of the *INS-IGF2* and *H19* regions between pig, human, and mouse show that the level of identity between human and pig is higher than that between human and mouse, or pig and mouse (Fig. 3). The exons are highly conserved, but a large number of intronic and intergenic regions are also remarkably well conserved among the three species. The Alfresco program (Jareborg and Durbin 2000) was used to perform a

comparative analysis between the pig and human sequences and to identify evolutionarily conserved elements (Fig. 1). The sequence of every conserved element was then compared to the mouse sequence by pairwise BLAST (http://www.ncbi.nlm.nih.gov) with the following parameters: word size 7; penalty for a mismatch –1. The elements that were found to be conserved also in mouse are indicated in Table 4. Alfresco was also used to compare the pig and mouse sequence and it gave similar results (data not shown). Several human regulatory elements already known were, as expected, found in the pig sequence, together with many other conserved elements whose function remains to be determined. Alfresco uses an algorithm named DBA (DNA Block Aligner; Jareborg et al. 1999), which was designed to identify small conserved motifs in non-coding sequences that are difficult to align. When comparing the human and pig sequences, we found that the results of DBA were identical to the results obtained with the pairwise BLAST by using defaults parameters (word size 11, penalty for a mismatch –2).

We have focused primarily on regulatory elements that are assumed to be involved in the regulation of *IGF2* expression. Two conserved elements upstream of the *insulin* promoter (no. 1 and 2 in Table 4) are probably involved in the regulation of transcription. There was no evidence a corresponding pig VNTR sequence as present about 500 bp upstream of the human *INS* gene (Bennett et al. 1995). The human VNTR consensus unit is ACAGGGGTGTGGGG which creates a very GC rich region with strong strand disequilibrium of G and C nucleotides. However, the two motifs AGGGG and TGGGG are found five and seven times, respectively, together with several similar motifs (for instance CGGGG, TGGGT, AGGGT, AGGGA, and AGGGC) in the corresponding pig sequence, but they are not organized in a tandem repeat. There is no strand disequilibrium in the pig because of the presence of several stretches of Cs.

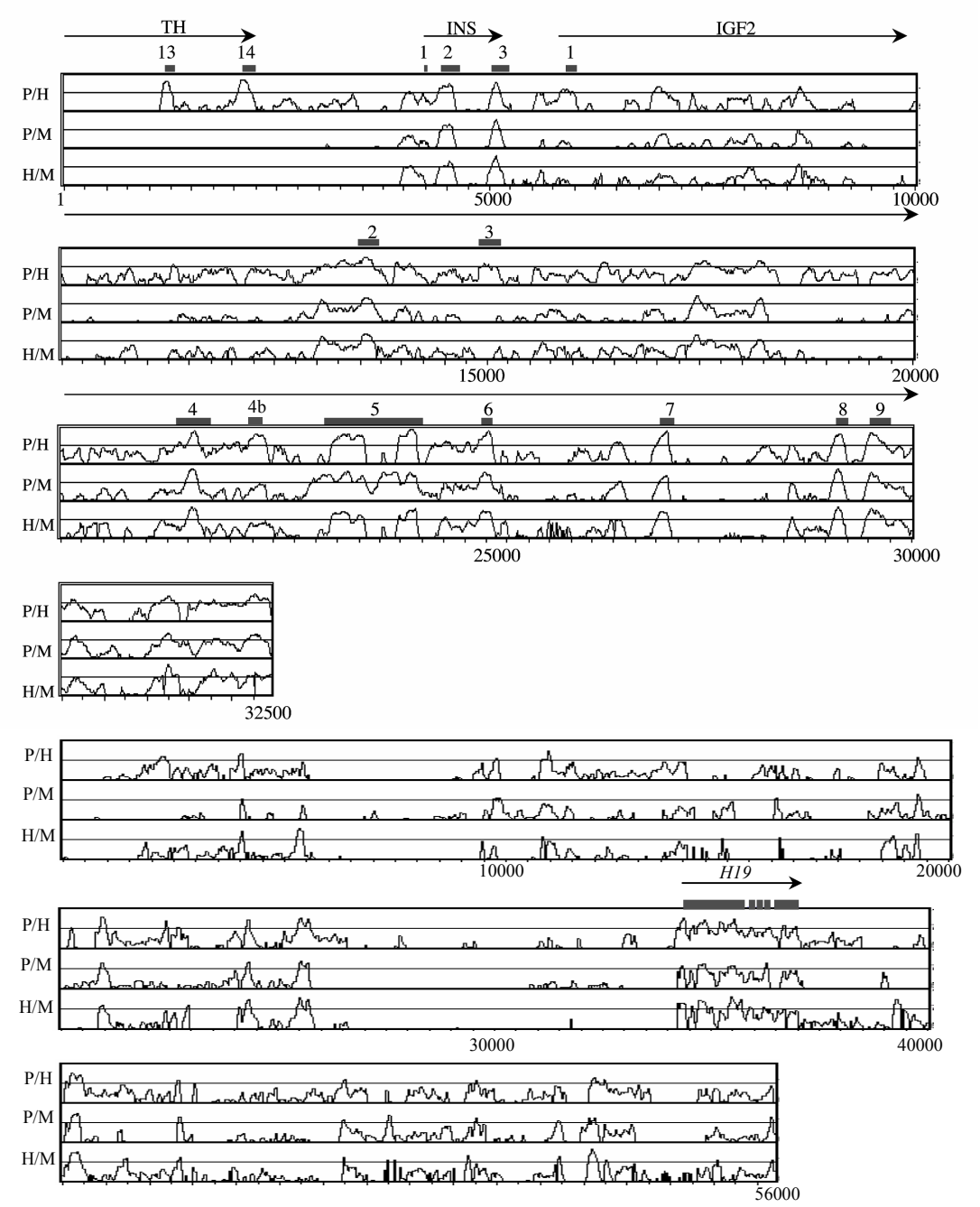


Fig. 3: Sequence identity plots for pairwise comparisons of the pig, human, and mouse *INS-IGF2*, and *H19* genomic regions. P/H: pig/human, P/M: pig/mouse, H/M: human/mouse (percent identity calculated with a window length of 100 bp). Conserved sequences are shown relative to their positions in the pig sequence (horizontal axes), and the percentage of identities (50%-100%) are indicated on the vertical axes. The locations of exons are shown above the profile.

A conserved CpG island of 650 bp is found about 3 kb upstream of *IGF2* exon 4 harboring 87 and 70 CpG dinucleotides in pig and human, respectively. Several short sequence elements within this CpG island are very well conserved (no. 27 to 31 in Table 4). This region

corresponds to a suggested control element, denoted differentially methylated region 1 (DMR1) in mouse, which shows a differential methylation pattern, and the unmethylated allele is associated with a silencer function in several mesodermal tissues (Constancia et al. 2000). The transcription factor GCF2 (GC binding factor) was suggested to repress the transcription through binding the unmethylated allele. Methylation would prevent binding of the repressor on the paternal allele (Eden et al. 2001). However, the imprinting status of *IGF2* in mouse skeletal muscle is independent of the DMR1 methylation suggesting an important, tissue specific regulation of the imprinting (Weber et al. 2001). The 54-bp core region of differentially methylated region 2 (DMR2) present in mouse *IGF2* exon 6 (Murrell et al. 2001) is also well conserved in the pig sequence (89% identity).

A large CpG island of about 3.5 kb containing the region from *IGF2* exon 4 to a part of intron 5 is found in both pigs and human. A large number of conserved sequences (no. 35 to 42 in Table 4) are found in this region and in a part of intron 5 outside the CpG island. Introns 4 and 5 harbor several highly conserved elements, which are likely to be involved in the regulation of transcription from promoters P3 and P4. Among these elements, the SP1 binding motif GGGGGCGGGGCGAGG upstream of the P3 promoter is perfectly conserved as well as the P3-4 element described by Rietveld et al. (1997). Many other conserved elements with a length varying from 17 to 158 bp and whose function is still unknown are spread all over the *INS-IGF2* region (Table 4). The *IGF2* 3' UTR harbors a large number of conserved elements. These include several simple repeats together with a group of non-repetitive elements (no. 49-59 in Table 4).

A complex pattern of conserved elements was found upstream of *H19*. A conserved motif repeated seven times in the human sequence was found three times in the pig (Fig. 4) between -1.1 and -2.6 kb from the transcription start of *H19*. These repeats, previously identified in human, mouse, and rat (Frevel et al. 1999; Bell and Felsenfeld 2000), contain a 12 bp consensus sequence which is a binding site for CTCF (CCCTC-binding factor). The sequence identity of the repeats between human and pig extends outside the 12 bp repeat in a motif covering 95 bp (Fig. 4). CTCF is a vertebrate regulation factor able to form a large variety of DNA – protein complexes involved in distinct functions including gene activation, repression, silencing, and chromatin insulation (Ohlsson et al. 2001). The 5' region of *H19* contains an imprinting mark characterized by paternal allele-specific methylation (Thorvaldsen et al. 1998). CTCF binding sites in this region are found in human, mouse, rat, and pig as well as in other imprinted domains having a similar regulation (Wylie et al. 2000) and were shown to

contribute to imprint regulation of neighboring genes by the formation of chromatin boundaries (Bell and Felsenfeld 2000, Kaffer et al. 2000).

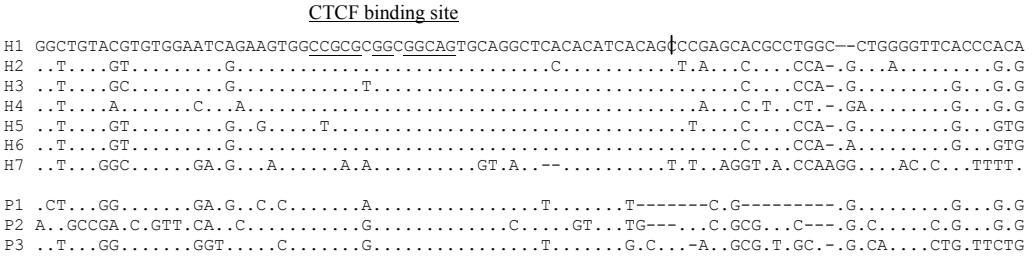


Fig. 4: Sequence alignment of the human (H1 to H7) and pig (P1 to P3) repeats upstream *H19* and containing the CTCF factor binding site. Identities to the human sequences are indicated by dots and alignment gaps are indicated by dashes.

Several enhancer elements previously characterized in human and/or mouse (Webber et al. 1998; Ishihara et al. 2000) were found upstream and downstream of *H19*. In the pig, the two endoderm-specific enhancers are situated at 3.1 and 5.6 kb downstream *H19* (no. 80 and 81 in Table 4). Four other conserved sequences (CS) identified by Ishihara et al. (2000) between mouse and human were found in the same region in pig (no. 82, 85-86-87, 89-90, and 91-92 in Table 4, similar to CS5, 6, 7, and 8, respectively). CS5, 6, and 7 exhibit a tissue-specific enhancer activity, and CS6 is particularly interesting because it is primarily active in skeletal muscle (Ishihara et al. 2000). We have also found several other elements that were previously identified to be conserved between human and mouse, but their functions are still unknown (no. 70, 72, and 94+95 in Table 4).

2.4. Discussion:

We present here a comparative sequence analysis of about 90 kb from the region containing the *INS – IGF2* and *H19* genes in pig, human, and mouse. The divergence between the human and pig lineage is estimated at 70 million years, whereas human and mice diverged approximately 100 million years ago (Andersson et al. 1996). The sequence similarity is high not only in exons but also in introns and intergenic regions. However, it is likely that the identity scores are somewhat inflated owing to the high GC content and the many CpG islands in this region. The sequence identity plots for the three pairwise species comparisons were remarkably similar (Fig. 3). The general trend was a higher sequence similarity between pig and human than between these species and mouse as expected from the phylogenetic relationship. There are some interesting discrepancies between species that may very well be functionally important. The intergenic distance between *TH* and *INS* is much larger in the

mouse (200 kb; Onyango et al. 2000) than it is in pig and human (2-3 kb). *IGF2* exon 3 is well conserved between pig and human (74%), but not in mouse (<50% identity compared with human or pig, Fig. 3). The 5' flanking region of *IGF2* exon 5 as well as part of this exon is well conserved between pig and mouse (>75%) but not in humans (<50%). These cases imply a faster substitution rate in one of these lineages suggesting an altered function for the actual region. The results illustrate how a comparative analysis based on three or more species add power to the interpretation of genome evolution.

The programs (DBA and BLAST) that were used to identify conserved elements between human and pig are designed to identify short conserved elements inside large genomic regions that cannot be aligned. The DBA algorithm (applied in Alfresco) is supposed to be more sensitive than BLAST when conserved motifs are very short (Jareborg et al. 1999). In our case, the short, highly conserved elements identified by BLAST and DBA correspond to the top of the peaks displayed on the sequence identity plots, and most of them are conserved between the three species. We identified several 17 bp elements that are 100% conserved between pig and human (Table 4). We believe that this approach is useful to detect short regulatory elements like transcription factor binding sites. A conserved sequence at a conserved position is likely to be functionally important. Other parameters were used by Onyango et al. (2000) when comparing the 1 Mb sequence covering the mouse and human imprinted domain. Their definition of a conserved elements was >100 bp with >70% nucleotide identity. All the elements we identified show >80% identity and several of them could be merged and considered as a single region with a lower degree of similarity, as illustrated on the identity plots. This would increase the proportion of aligned sequences (outside genes and promoters), in our case 9.7 and 4.6% in the *INS-IGF2* and *H19* regions, respectively, to a level closer to the 35.8% estimated for the human/mouse comparison of the *H19* and *Mash2* region. (Onyango et al. 2000) When the same programs and same parameters were used to compare the human *H19* region with the pig *INS-IGF2* region, no conserved sequences were identified between these two unrelated regions. This shows that the majority of the conserved regions reported in this study are not random sequence similarities.

Our sequence data revealed the exon/intron organization, promoters, and other potential regulatory regions of the *INS*, *IGF2*, and *H19* genes in pig. The three genes are very well conserved between human and pig. The ten exons of human *IGF2* and the four promoters were identified in pig by comparative sequencing. Our RT-PCR analysis confirmed that all ten *IGF2* exons and the four promoters are used in the pig. It is worth noticing that the number of *IGF2* exons in humans are generally considered to be nine but our observation of a

minor transcript containing a tenth exon (denoted 4b) is in perfect agreement with Mineo et al. (2000).

A very high GC content and an exceptional concentration of CpG islands characterize this genomic region in both human and pig. This tendency is significantly stronger in pig than in human, in which it was already shown to be higher than in mouse (Onyango et al. 2000). The small amount of pig genomic sequence publicly available is too limited to know whether this higher GC content is a general trend of the pig genome compare with the human genome or if this is specific to this region. However, an independent comparative study of a 130 kb genomic region on pig Chr 15 and human Chr 2 (V. Amarger. in preparation) shows a similar GC content in both species (45.6 and 45.4 % in pig and human, respectively). The presence and importance of CpG islands in the vicinity of imprinted genes is now well established. Although their action is not clearly demonstrated, it appears that CpG islands are necessary to establish and maintain imprinting patterns (Paulsen et al. 2000, Engemann et al. 2000). The role of CpG islands in gene regulation is strongly supported by comparison of the *Impact* gene that is imprinted in mouse but not in humans. The two homologs display a striking difference in the CpG island pattern in their upstream promoter region (Okamura et al. 2000). CpG islands are often associated with imprinting control regions (ICR) harboring a differential allelic methylation. The way methylation of these ICRs regulates expression involves methylation sensitive DNA binding factors that can act as insulators (Holmgren et al. 2001) or repressors (Eden et al. 2001).

Another striking feature is the low proportion of interspersed repeats in the close vicinity of *INS*, *IGF2*, and *H19*, particularly in the pig (Table 1). It is possible that the pig estimates are slightly biased, since interspersed repeats are more studied in human but this cannot explain the observation of the large difference in interspersed repeats between the pig and human *H19* regions, 1.7% vs 8.2%. Furthermore, this difference is accompanied by a similar difference in GC content that should be unbiased. This remarkably low proportion of interspersed repeats seems to be specific to the *INS-IGF2-H19* region since 30% of the human 1Mb region on chr11p15 is composed of interspersed repeats (Onyango et al. 2000), and the average frequency for the human genome is about 45% (Lander et al. 2001). Thus, the *INS-IGF2* region in pigs and human and the *H19* region in pigs is as devoid of interspersed repeats as the HOX gene clusters which recently were identified as being some of the most repeat-poor regions in the human genome (Lander et al. 2001). A low amount of interspersed repeats was also observed in the imprinted domain spanning from the *KCNQ1* to *CARS* genes together with a remarkably uneven distribution of these repeats, which appears to be conserved

between human and mouse (Engemann et al. 2000). The reason why very few interspersed repeats are present in the *INS-IGF2-H19* genes is not clear. However, the phenomenon may be related to the complicated regulation of gene expression and imprinting of these genes and introduction of foreign sequences may disturb essential *cis*-regulatory mechanisms. In contrast, the frequency of simple repeats in the *INS/IGF2* region is slightly higher than the genome average (~4% vs. ~3%). It has previously been suggested that simple repeats may play a role in the regulation of imprinted genes (Shibata et al. 1998) but no clear mechanism has been characterized so far. Among these repeats, a complex microsatellite element (several CA stretches with interruptions) in the 3' UTR of *IGF2* is conserved between human, pig, horse, and mouse (Jeon et al. 1999).

The VNTR upstream of the human insulin gene evidently has a functional role but it was not found in the pig sequence. However, several 5 bp motifs present in the VNTR unit were found interspersed in the corresponding region in the pig. It is not clear how this VNTR affects insulin expression, but it may influence the chromatin structure, modulating the accessibility of transcription factors. It could also be a transcription factor-binding site. The fact that this VNTR is not conserved between two quite closely related species like human and pig, while many other regulatory elements are very well conserved, suggests that the function of this sequence might not be directly related to its repetitive structure. VNTRs are predominantly present in subtelomeric regions in the human genome (Amarger et al. 1998).

Several repeats are present in the 5' region of *H19*. This region has been shown to contain an epigenetic mark required for the imprinting of both *H19* and *IGF2* (Tremblay et al. 1995; Thorvaldsen et al. 1998). It acts as an insulator (Kaffer et al. 2000), and its activity is dependent upon the vertebrate enhancer-blocking protein CTCF (Bell and Felsenfeld 2000). The methylation status of this region seems to be the major mechanism by which the insulator activity is modulated (Holmgren et al. 2001, Reed et al. 2001), and CTCF function marks the *IGF2/H19* expression domain in a parent-of-origin-dependent manner (Ohlsson et al. 2001). The loss of imprinting of *IGF2* correlated with biallelic hypermethylation of this region was suggested to give a predisposition to colorectal cancer (Nakagawa et al. 2001). However, aberrant methylation is not always associated with abnormal imprinting (Cui et al. 2001). In the pig, three copies of this binding site are found in a conserved larger repeated element situated –1.1 to –2.6 kb upstream of *H19*. A silencer element involved in the regulation of the imprinting of the paternal allele in a methylation-insensitive manner overlaps this insulator region in the mouse (Drewell et al. 2000: Ferguson-Smith 2000). However, we did not find any evidence of conservation of this region in the pig. Furthermore, a skeletal muscle-specific

element involved in the silencing of the *IGF2* maternal allele was identified in the murine *IGF2-H19* intergenic region (Ainscough et al. 2000), but our pig sequence does not completely cover this region. Because of its tissue-specific function, this element is of potential interest for further studies of the *IGF2*-linked QTL in the pig. The 3' region of *H19* harbors several enhancer elements that affect expression of both *IGF2* and *H19* (Webber et al. 1998). Besides the two well characterized endoderm-specific enhancers, other putative enhancer elements that could have a tissue-specific action are present in this region.

This study revealed a large number of conserved elements that might have a functional role in regulating *IGF2* expression. Several approaches can now be used to identify the molecular basis for the paternally expressed QTL affecting muscle development in the pig. From the sequence information presented here, a number of genetic markers can be developed and used to define haplotypes associated with different QTL alleles. This strategy would allow us to narrow the candidate region. A search for polymorphisms associated with the phenotype would then be conducted among the conserved elements in the defined region. However, the results of the present study shows that this will still be a challenge owing to the large number of conserved elements potentially influencing *IGF2* function, and we cannot exclude the possibility that some QTL alleles may be due to the combined effect of multiple substitutions. An important topic for future research is also to study *IGF2* expression, in particular in skeletal muscle. The sequence information provided here will facilitate these studies.

Acknowledgments

This work was supported by the Swedish Research Council for Forestry and Agriculture, Swedish Foundation for Strategic Research, Seghers Genetics, and a grant from the Belgian Ministry of Agriculture (D1/2 – 5795A). The authors thank R. Erlandsson and B. Amini for sequencing a part of the BAC370 at the Genome center, Royal Institute of Technology, Stockholm, Niclas Jareborg for his help with Alfresco, Erik Bongcam-Rudloff for bioinformatic assistance, and Göran Andersson for comments on the paper.

CHAPTER 3

A REGULATORY MUTATION IN IGF2 CAUSES A MAJOR QTL EFFECT ON MUSCLE GROWTH IN THE PIG

Anne-Sophie Van Laere[1]*, Minh Nguyen[3]*, Martin Braunschweig[1], Carine Nezer[3], Catherine Collette[3], Laurence Moreau[3], Alan L. Archibald[4], Chris S. Haley[4], Nadine Buys[5], Michael Tally[6], Go¨ran Andersson[1], Michel Georges[3] and Leif Andersson[1,2]

Nature, Volume 425: 832-836, 2003

[1]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, and

[2]Department of Medical Biochemistry and Microbiology, Uppsala University, BMC, Box 597, SE-751 24 Uppsala, Sweden

[3]Department of Genetics, Faculty of Veterinary Medicine, University of Liège (B43), 20, bd. de Colonster, 4000 Liège, Belgium

[4]Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, Scotland, UK

[5]Gentec, Kapelbaan 15, 9255 Buggenhout, Belgium

[6]Tally Consulting, SE-11458 Stockholm, Sweden

* These authors contributed equally to this work

Minh Nguyen's contribution:

1. DNA resequencing and sequence analyses of eight distinct haplotypes for the 28.6 kb fragment
2. DNA methylation analysis using bisulphite sequencing
3. Imprinting analysis of *IGF2* in fetal and postnatal porcine skeletal muscle

Summary

Most traits and disorders have a multifactorial background indicating that they are controlled by environmental factors as well as an unknown number of quantitative trait loci (QTLs) (Mackay, 2001; Andersson, 2001). The identification of mutations underlying QTLs is a challenge because each locus explains only a fraction of the phenotypic variation (Glazier et al., 2002; Darvasi and Pisante-Shalom, 2002). A paternally expressed QTL affecting muscle growth, fat deposition and size of the heart in pigs maps to the *IGF2* (insulin-like growth factor 2) region (Jeon et al., 1999; Nezer et al., 1999). Here we show that this QTL is caused by a nucleotide substitution in intron 3 of *IGF2*. The mutation occurs in an evolutionarily conserved CpG island that is hypomethylated in skeletal muscle. The mutation abrogates in vitro interaction with a nuclear factor, probably a repressor, and pigs inheriting the mutation from their sire have a threefold increase in *IGF2* messenger RNA expression in postnatal muscle. Our study establishes a causal relationship between a single-base-pair substitution in a non-coding region and a QTL effect. The result supports the long-held view that regulatory mutations are important for controlling phenotypic variation (King and Wilson, 1975).

3.1. Introduction:

The QTL affecting muscle growth, fat deposition and heart size was first identified in intercrosses between the European wild boar and Large White domestic pigs and between Piétrain and Large White pigs (Jeon et al., 1999; Nezer et al., 1999). The alleles from the Large White breed in the first cross and the Piétrain breed in the second cross increased muscle mass and reduced back-fat thickness. The QTL explained 15–30% of the phenotypic variation in muscle mass and 10–20% of the variation in back-fat thickness (Jeon et al., 1999; Nezer et al., 1999). We recently used a haplotypes sharing approach to refine the map position of the QTL (Nezer et al., 2003). We assumed that a new allele (Q) promoting muscle development occurred g generations ago on a chromosome carrying the wildtype allele (q). We also assumed that the favorable allele had gone through a selective sweep due to the strong selection for lean growth in commercial pig populations. A QTL genotype cannot be deduced directly from an individual's phenotype but the QTL genotype of sires can be determined by progeny testing and marker-assisted segregation analysis (Andersson, 2001). Twenty-eight chromosomes with known QTL status were identified. All 19 Q-bearing chromosomes shared a haplotype in the 250-kilobase (kb) interval between the markers 370SNP6/15 and SWC9 (*IGF2* 3' untranslated region), which was therefore predicted to contain the QTL. This region contains *INS* and *IGF2* as the only known paternally expressed

genes. Given their known functions and especially the role of IGF2 in myogenesis (Florini et al., 1995), they stood out as prime positional candidates.

3.2. Results and Discussion:

We re-sequenced one of the 19 Q chromosomes (P208) and six q chromosomes (each corresponding to a distinct marker haplotype) for a 28.6-kb segment containing *IGF2*, *INS* and the 3' end of *TH*. This chromosome collection was expanded by including Q and q chromosomes from the following: (1) a wild boar/Large White intercross segregating for the QTL (Jeon et al., 1999); (2) a Swedish Landrace boar showing no evidence for QTL segregation in a previous study (Evans et al., 2003); and (3) F1 sires from a Hampshire/Landrace cross and a Meishan/Large White intercross both showing no indication for QTL segregation (see Methods). The lack of evidence for QTL segregation shows that the boars are either homozygous Q/Q or q/q. A Japanese wild boar was included as a reference for the phylogenetic analysis and it was assumed to be homozygous wild type (q/q). We identified a total of 258 DNA sequence polymorphisms corresponding to one polymorphic nucleotide per 111 base pairs (bp) (Fig. 1). Two major and quite divergent clusters of haplotypes were revealed (Supplementary Fig. 1). The two established Q haplotypes from Piétrain and Large White animals (P208 and LW3) were identical to each other and to the chromosomes from the Landrace (LRJ) and Hampshire/ Landrace (H205) animals, showing that the latter two must be of Q type as well. The absence of QTL segregation in the offspring of the F1 Hampshire x Landrace boar carrying the H205 and H254 chromosomes implies that the latter recombinant chromosome is also of Q type. This places the causative mutation downstream from *IGF2* intron 1, in the region for which H254 is identical to the other Q chromosomes. The Large White chromosome (LW197) from the Meishan/Large White pedigree clearly clustered with q chromosomes, implying that the F1 sire used for sequencing was homozygous q/q as no overall evidence for QTL segregation was observed in this intercross. This is consistent with a previous study showing that Meishan pigs carry an *IGF2* allele associated with low muscle mass (de Koning et al., 2000). Surprisingly, the Meishan allele (M220) was nearly identical to the Q chromosomes but with one notable exception, it shared guanine with all q chromosomes at a position (*IGF2*-intron3-nucleotide 3072) where all Q chromosomes have adenine (Fig. 1).

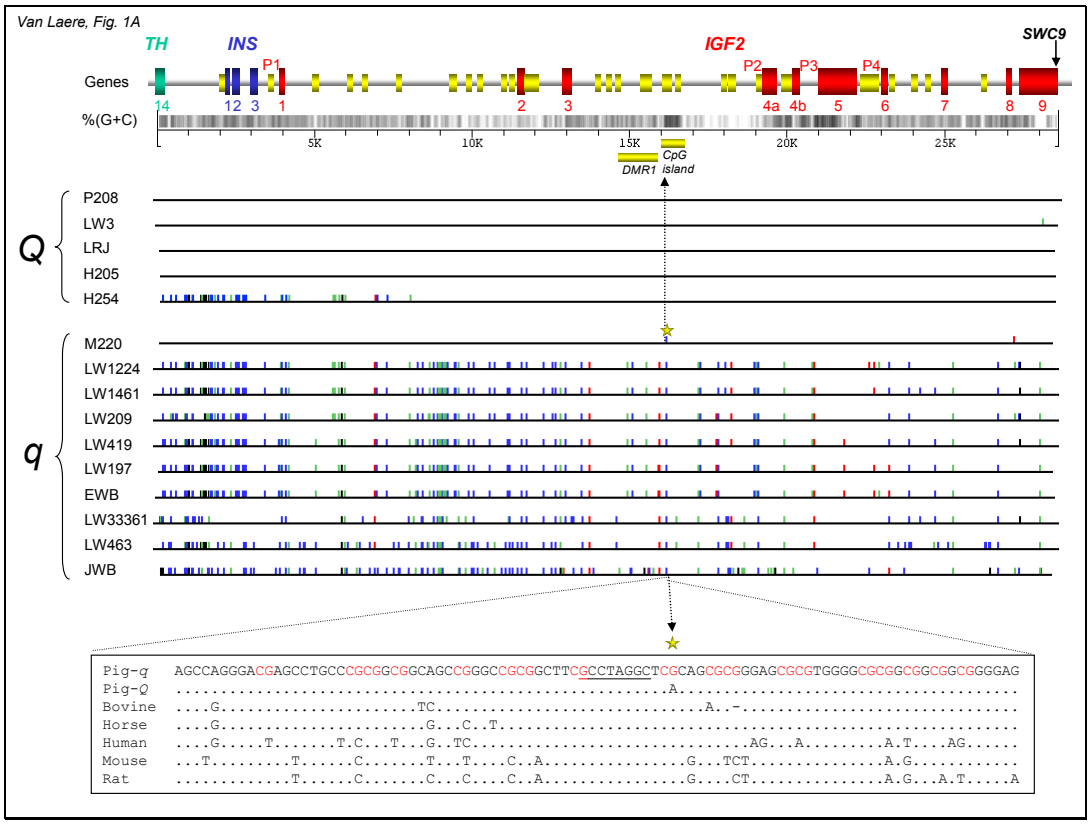


Figure 1: Polymorphisms in a 28.6-kb segment containing *TH* (exon 14), *INS* and *IGF2* among 15 pig chromosomes with deduced QTL status. The (G+C) content of a moving 100-bp window is shown on a grey scale (black 100%, white 0%). Yellow cylinders mark evolutionarily conserved regions (Amarger et al., 2002). Viewgene (Kashuk et al., 2002) was used to highlight differences between the reference P208 sequence and other chromosomes. Blocks indicate transitions (blue), transversions (green), insertions (red) and deletions (black). The QTN is indicated with an asterisk and the surrounding sequence is shown for eight mammals (CpGs are highlighted in red; a palindromic sequence is underlined). P, Piétrain; LW, Large White; LR, Landrace; H, Hampshire; M, Meishan; EWB, European wild boar; JWB, Japanese wild boar.

Under a bi-allelic QTL model, the causative mutation would correspond to a DNA polymorphism for which the two alleles segregate perfectly between Q and q chromosomes. The G to A transition at *IGF2*-intron3-3072 is the only polymorphism fulfilling this criterion, implying that it is the causative quantitative trait nucleotide (QTN) (Mackay, 2001). We genotyped the founders and 12 F1 boars for the putative QTN to verify the QTL status of animals from the Meishan/Large White intercross. All founders and F1 boars were homozygous G/G (q/q) except one Large White founder and one F1 boar, which were heterozygous A/G. A QTL analysis revealed that the heterozygous boar, but no other F1 sire, showed clear evidence for segregation of a paternally expressed QTL, and the Meishan allele increased back-fat thickness as predicted (X^2 with 1 degree of freedom = 7.75, P = 0.005). Including this, we have so far tested 13 large sire families where the sire is heterozygous A/G

at the QTN, and all have shown evidence for QTL segregation. In contrast, we have tested more than 50 sires, representing several different breeds, genotyped as homozygous A/A or G/G without obtaining any evidence for the segregation of a paternally expressed QTL at the *IGF2* locus. The results provide conclusive genetic evidence that *IGF2*-intron3-G3072A is the causative mutation. The Meishan allele is apparently identical to the ancestral haplotype on which the mutation occurred.

IGF2-intron3-3072 is part of an evolutionarily conserved CpG island of unknown function (Amarger et al., 2002) located between differentially methylated region 1 (DMR1) and a matrix attachment region previously defined in mice (Greally, 1997; Constancia et al., 2000; Eden et al., 2001). The 94-bp sequence around the mutation shows about 85% identity to human, and the wild-type nucleotide at *IGF2*-intron3-3072 is conserved among eight mammalian species (Fig. 1). The QTN occurs 3 bp downstream of a conserved 8-bp palindrome. The methylation status of the 300-bp fragment centered on *IGF2*-intron3-3072 and containing 50 CpG dinucleotides was examined by bisulphite sequencing in four-month-old Qpat/qmat and qpat/Qmat pigs. The CpG island was methylated in liver (26% of CpGs methylated on average), whereas in skeletal muscle both paternal (pat) and maternal (mat) chromosomes were essentially unmethylated (including the *IGF2*-intron3-3071 C residue) irrespective of QTL genotype (3.4% of CpGs methylated on average, Fig. 2a; see also Supplementary Fig. 2).

To uncover a possible function for this element, we performed electrophoretic mobility shift assays (EMSAs) using wild-type (q) and mutant (Q) sequences. Nuclear extracts were incubated with radioactively labeled q or Q double-stranded oligonucleotides. One specific complex (C1 in Fig. 2b) was obtained with the unmethylated wild-type (q) but not the mutant (Q) probe using extracts from murine C2C12 myoblasts. This complex was not obtained with the q* probe, which has a methylated CpG at the QTN (Fig. 2b). A complex with approximately the same migration—but slightly weaker—was also detected in extracts from human HEK293 embryonic kidney cells and HepG2 hepatocytes. The specificity of the complex was confirmed as competition was obtained with 10- fold molar excess of unlabelled q probe, whereas a 50-fold excess of unlabelled Q probe or methylated q* probe did not compete (Fig. 2b). Thus, the wild-type sequence binds a nuclear factor, and this interaction is abrogated by the mutation or methylation of the actual CpG site.

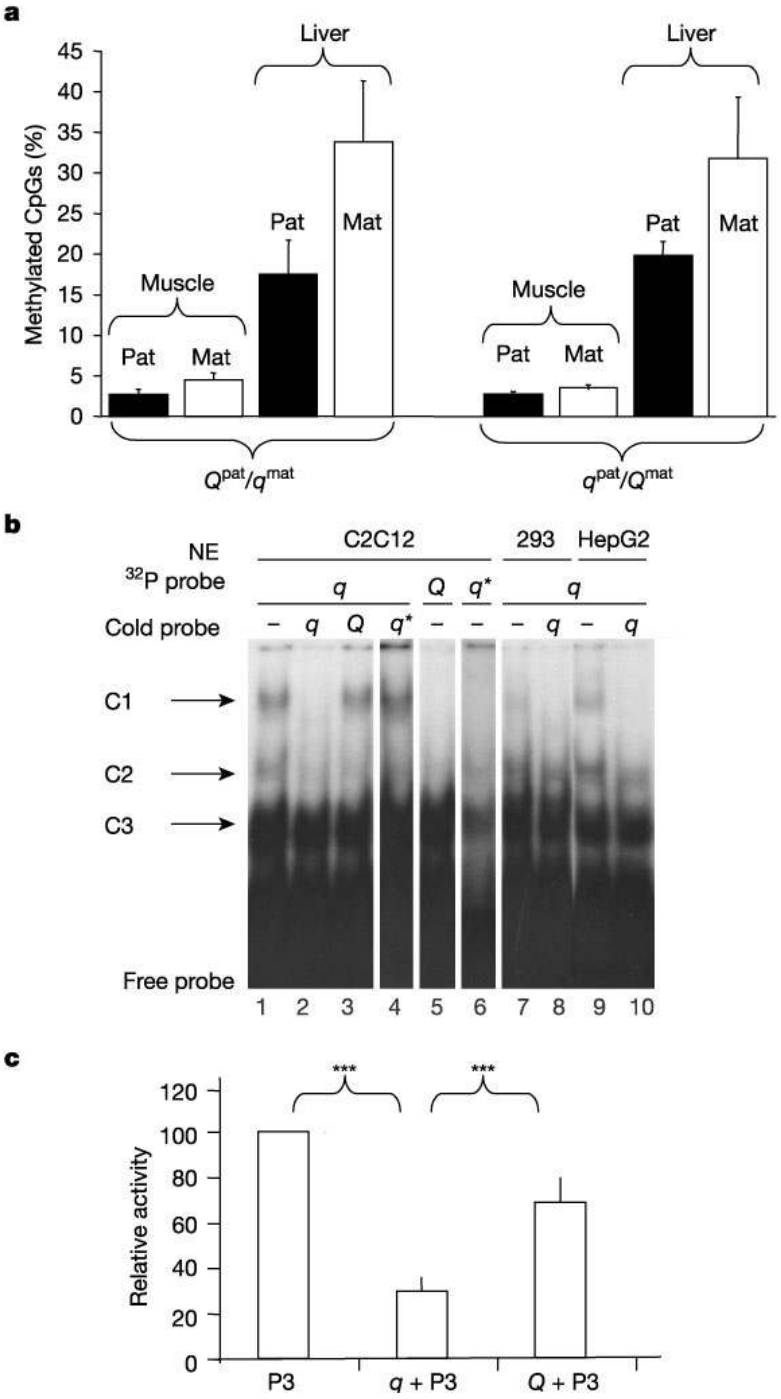


Figure 2: Methylation status, EMSA and transfection assays assessing the significance of the *IGF2* mutation. a, Percentage methylation around the QTN in liver and skeletal muscle of four-month-old pigs. Means ± s.e.m. are given; the numbers of analyzed paternal (Pat) and maternal (Mat) chromosomes were in the range 10–26. b, EMSA using nuclear extracts (NE) from C2C12, HEK293 and HepG2 cells. Complex 1 (C1) was exclusively detected with the q probe. C2 was stronger in q but also probably present in the Q lane. C3 was unspecific. c, Luciferase assays of reporter constructs using pig *IGF2* P3 promoter and intron 3 fragments (Q and q). Relative activities compared with the P3–LUC reporter are reported (means ± s.e.m.). Triple asterisk, P<0.001.

We analyzed the effect of the *IGF2* Q mutation on transcription by using a transient transfection assay in mouse C2C12 myoblasts. We made Q and q constructs containing a 578-bp fragment from the actual region inserted in front of a luciferase reporter gene driven by the endogenous pig *IGF2* promoter 3 (P3) located approximately 5 kb downstream from the QTN (Fig. 1). This promoter was chosen because it generates the predominant *IGF2* transcript in muscle and the QTN affected the amount of P3 transcripts in vivo (see below). The q fragment clearly acted as a repressor element and reduced luciferase activity to about 25%, whereas the Q fragment was a significantly weaker repressor element and showed about 70% activity compared with P3 alone (Fig. 2c). Our interpretation of this result, combined with those from the EMSA experiment, is that the Q mutation abrogates the interaction with a putative repressor protein. Thus, the *IGF2*-intron3-G3072A transition may be sufficient to explain the QTL effect as the two constructs only differ at this position. The constructs were also inserted in front of the heterologous herpes *thymidine kinase* (*TK*) minimal promoter. In this case the q construct caused a twofold increase of transcription whereas the Q construct caused a significantly higher, sevenfold increase (Supplementary Fig. 3). The results support our interpretation that the q allele represses transcriptional activity in C2C12 myoblasts when compared with the Q allele.

The in vivo effect of the mutation on *IGF2* expression was studied in a purpose-built Q/q x Q/q intercross. We tested the effect of the intron3-3072 mutation on *IGF2* imprinting, as a deletion encompassing DMR0, DMR1 and the associated CpG island derepresses the maternal *IGF2* allele in mesodermal tissues in mouse (Constancia et al., 2000). This was achieved by monitoring transcription from paternal and maternal alleles in tissues of q/q, Qpat/qmat and qpat/Qmat animals heterozygous for the SWC9 microsatellite located in the *IGF2* 3' untranslated region. Before birth, *IGF2* was expressed exclusively from the paternal allele in skeletal muscle and kidney, irrespective of QTL genotype. At four months of age, weak expression from the maternal allele was observed in skeletal muscle, however at comparable rates for all three QTL genotypes (Supplementary Fig. 4). Only the paternal allele could be detected in four-month-old kidney. Consequently, the mutation does not seem to affect the imprinting status of *IGF2*. The partial derepression of the maternal allele in skeletal muscle may however explain why in a previous study muscle growth was found to be slightly superior in qpat/Qmat versus q/q animals, and in Q/Q versus Qpat/qmat animals (Jeon et al., 1999).

The Q allele was expected to be associated with increased *IGF2* expression because IGF2 stimulates myogenesis (Florini et al., 1995). We monitored the relative mRNA expression of

IGF2 at different ages in the Q/q x Q/q intercross using both northern blot analysis and real-time polymerase chain reaction (PCR) (Fig. 3). The expression levels in fetal muscle and postnatal liver at three weeks of age were approximately tenfold higher compared with postnatal muscle. No significant difference was observed in fetal samples or in postnatal liver samples, but a significant threefold increase of postnatal *IGF2* mRNA expression in skeletal muscle was observed in Q/Q or Qpat/qmat versus qpat/Qmat or q/q progeny. There was also a significant, but less pronounced, increase of mRNA expression in heart associated with the Qpat allele. The significant difference in *IGF2* expression revealed by real-time PCR was confirmed using two different internal controls: *GAPDH* (Fig. 3b) and *HPRT* (data not shown). We found an increase of all detected transcripts originating from the three promoters (P2–P4) located downstream of the QTN. The results provide strong support for *IGF2* being the causative gene. The significant differences in *IGF2* mRNA expression between genotypes in skeletal and cardiac muscle and the lack of significant differences in fetal muscle and postnatal liver are consistent with our previous data showing clear phenotypic effects of the *IGF2* QTL on muscle growth and size of the heart but no effects on birth weight or weight of liver (Jeon et al., 1999). The results demonstrate that IGF2 has an important role in regulating postnatal myogenesis. The higher expression in postnatal skeletal and cardiac muscle associated with the Q allele parallels the observation of a continued postnatal *IGF2* expression in mesodermal tissue of transgenic mice carrying a 5-kb deletion of *IGF2* encompassing DMR1 and the associated CpG island (Constancia et al., 2000).

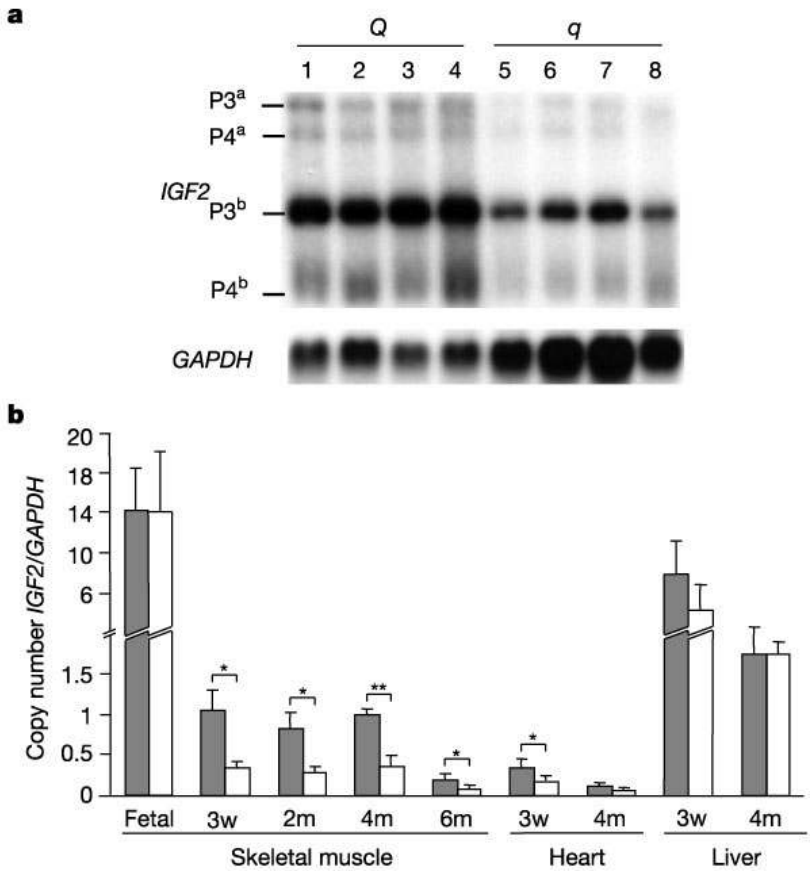


Figure 3: Analysis of *IGF2* mRNA expression. a, Northern blot of skeletal muscle (gluteus) poly(A)þ RNA from three-week-old piglets. Animals 1–4 and 5–8 carried a paternal *IGF2**Q or *q allele, respectively. P3 and P4 indicate promoter usage, and a/b superscripts indicate the alternative polyadenylation signal used. All four transcripts showed a higher relative expression (standardized using *GAPDH*) in the *Q group ($P<0.05$). b, Real-time PCR analysis of *IGF2* expression in pigs carrying paternal *IGF2**Q (grey columns) or *q (white columns) alleles. Expression levels were normalized using *GAPDH*. Means = s.e.m. are given, n = 3–11. Asterisk, $P < 0.05$; double asterisk, $P < 0.01$; w, week; m, month.

Immunoreactive serum levels of IGF2 were determined by a radioimmunoassay, but no significant differences between genotypes were observed (Supplementary Fig. 5). This finding was expected because the major source of IGF2 in serum is generated from liver, where no difference in IGF2 expression was detected (Fig. 3b). The results suggest that locally produced IGF2 in muscle determines the phenotype, as also indicated by the observation that there is no general overgrowth in pigs expressing the Q allele but rather a changed body composition.

There has been a strong selection for lean growth (high muscle mass and low fat content) in commercial pig populations over the past 50 yr. Therefore we investigated how this selection pressure has affected the allele frequency distribution of the *IGF2* QTL. The causative mutation was absent in a small sample of European and Asian wild boars and in breeds that

have not been strongly selected for lean growth (Supplementary Table 1). In contrast, the causative mutation was found at high frequencies in several breeds that have been subjected to strong selection for lean growth. This confirms our prior assumption that *IGF2**Q has experienced a selective sweep and has been spread between breeds by cross-breeding. The *IGF2**Q mutation increases meat production, at the expense of fat, by 3–4%. The high frequency of *IGF2**Q among major pig breeds implies that this mutation has had a large impact on pig production. European and Asian pigs were domesticated from different subspecies of the wild boar, and Asian germplasm was introgressed into European pig breeds during the eighteenth and nineteenth centuries (Giuffra et al., 2000). The *IGF2**Q mutation apparently occurred on an Asian chromosome as it shows a very close relationship to the haplotype carried by Chinese Meishan pigs. This explains the large genetic distance between Q and q haplotypes present in European domestic pigs (Fig. 1; see also Supplementary Fig. 1).

We have achieved an extraordinary resolution, down to a single nucleotide difference, in the genetic analysis of a QTL. This was possible by exploiting a selective sweep of a favorable mutation in commercial pig populations. Determination of complete genome sequences from major farm animals in the near future will allow the exploitation of the full potential of farm animal genomics. High density single-nucleotide polymorphism typing of domestic animals will allow identification of selective sweeps as documented here. Furthermore, re-sequencing of the entire genome using samples from different breeds, including wild ancestral species, will be very fruitful for studying genotype–phenotype relationships. This is well illustrated by the recent identification of the causative mutations for several interesting phenotypes in domestic animals (Grobet et al., 1997; Milan et al., 2000; Galloway et al., 2000; Mulsant et al., 2001; Pailhoux et al., 2001; Freking et al., 2002; Grisart et al., 2002). For instance, the callipyge mutation in sheep shares several features with the *IGF2* mutation in pigs; it affects body composition (muscle/fat content), shows a parent-of-origin effect and is a single-base substitution in a non-coding region (Freking et al., 2002). In fact, farm animal genetics provides major advantages compared with human genetics, and in some aspects compared with model organisms, for genetic dissection of multifactorial traits (Andersson, 2001). Farm animals are emerging as prime model organisms for understanding the genetic basis for multifactorial traits.

3.3. Materials and Methods:

3.3.1. Marker-assisted segregation analysis:

QTL genotyping of the Piétrain/Large White, wild boar/Large White and Hampshire/ Landrace crosses was performed as described (Nezer et al., 2003). Briefly, the likelihood of the pedigree data was computed under two hypotheses: H0, postulating that the corresponding boar was homozygous at the QTL (Q/Q or q/q), and H1, postulating that the boar was heterozygous Q/q. Likelihoods were computed using 'percentage lean meat' as phenotype, and assuming an allele substitution effect of 3.0% (Nezer et al., 1999). If the probability in favor of one of the hypotheses were superior or equal to 100:1, the most likely hypothesis was considered to be true. The Meishan/Large White cross consisted of 703 F2 animals with data on back-fat depths. An interval analysis approach (Haley et al., 1994) with microsatellite markers spanning the *IGF2* region revealed no indication of an overall QTL effect, imprinted or not.

3.3.2. DNA sequencing:

Animals homozygous for 13 of the haplotypes of interest were identified using flanking markers and pedigree information. A 28.6-kb segment was amplified from genomic DNA in seven long-range PCR products using the Expand Long Template PCR system (Roche Diagnostics GmbH). The same procedure was used to amplify the remaining M220 and LW197 haplotypes from two BAC clones isolated from a genomic library made from a Meishan/Large White F1 individual (Anderson et al., 2000). PCR products were purified using Geneclean (Polylab) and sequenced using the Big Dye Terminator Sequencing or dGTP Big Dye Terminator kits (Perkin Elmer). All primers are given in Supplementary Table 2. The sequence traces were assembled and analyzed for DNA sequence polymorphism using Polyphred/Phrap/Consed (Nickerson et al., 1997).

3.3.3. Genotyping of IGF2-intron3-3072:

Genotyping was done by pyrosequencing (Pyrosequencing AB). A 231-bp DNA fragment was PCR-amplified using Hot Star Taq DNA polymerase and Q-Solution (Qiagen) with the primers 18274F (5'-biotin-GGGCCGCGGCTTCGCCTAG-3') and 18274R (5'-CGCAC-GCTTCTCCTGCCACTG-30). The sequencing primer (5'-CCCCACGCGCTCCCG-CGCT-3') was designed on the reverse strand because of the palindrome located 5' of the QTN.

3.3.4. Bisulphite-based methylation analysis:

Bisulphite sequencing was performed as described (Engemann et al., 2002). A 300-bp fragment centred around intron3-3072 was amplified using a two-step PCR reaction with the following primers: PCR1-UP, 5'-TTGAGTGGGGATTGTTGAAGTTTT-3'; PCR1-DN, 5'-ACCCACTTATAATCTAAAAAAATAATAAATATATCTAA-3'; PCR2-UP, 5'-GGGGA-TTGTTGAAGTTTT-3'; PCR2-DN, 5'-CTTCTCCTACCACTAAAAA-3'. The amplified strand was chosen in order to differentiate the Q and q alleles. PCR products were cloned in the pCR2.1 vector (Invitrogen). Plasmid DNA was purified (Plasmid mini kit, Qiagen) and sequenced (Big Dye Terminator kit, Perkin Elmer). Inserts with identical sequences, that is having the same combination of C residues (whether part of a CpG dinucleotide or not) converted to U, were considered to derive from the same PCR product and were only considered once.

3.3.5. Electrophoretic mobility shift assays:

DNA-binding proteins were extracted as described (Andrews and Faller, 1991). EMSAs were performed with 40 fm 32P-labelled, double-stranded oligonucleotide, 10 mg nuclear extract and 2 mg poly(dI-dC) in binding buffer (15 mM HEPES pH 7.65, 30.1 mM KCl, 2 mM MgCl2, 2 mM spermidine, 0.1 mM EDTA, 0.63 mM dithiothreitol, 0.06% NP-40, 7.5% glycerol). For competition assays a 10-, 20-, 50- and 100-fold molar excess of cold double-stranded oligonucleotide were added. Reactions were incubated for 20 min on ice before ^{32}P-labelled, double stranded oligonucleotide was added. Binding was allowed to proceed for 30 min at room temperature. DNA–protein complexes were resolved on a 5% native polyacrylamide gel run in TBE x0.5 at room temperature for 2 h at 150 V. The following two (Q/q) 27-bp unmethylated oligonucleotides were used: 50-GATCCTTCGCCTAGGCTC-(A/G)CAGCGCGGGAGCGA-3'. A methylated q probe (q*) was generated by incorporating a methylated cytosine at the mutated CpG site during oligonucleotide synthesis.

3.3.6. Transient transfection assay:

The constructs contained 578 bp from *IGF2* intron 3 (nucleotides 2868–3446), followed by the *IGF2* P3 promoter (nucleotides - 222 to + 45 relative to the start of transcription) (Amarger et al., 2002) and a luciferase reporter. C2C12 myoblast cells were grown to approximately 80% confluence. Cells were transiently co-transfected with the firefly luciferase reporter construct (4 mg) and a Renilla luciferase control vector (phRG-TK, Promega; 80 ng) using 10 mg Lipofectamine 2000 (Invitrogen). Cells were incubated for 25 h

before lysis in 100 ml Triton lysis solution. Luciferase activities were measured using the Dual-Luciferase Reporter Assay System (Promega). The results are based on four triplicate experiments using two independent plasmid preparations for each construct. Statistical analysis was done with an analysis of variance.

3.3.7. Northern blot analysis and real-time RT–PCR:

Total RNA was prepared using Trizol (Invitrogen) and treated with DNase I (Ambion). Products from the first-strand complementary DNA synthesis (Amersham Biosciences) were purified with QIAquick columns (Qiagen). Poly(A)$^+$ RNA was then isolated using the Oligotex mRNA kit (Qiagen). Poly(A)$^+$ mRNA (about 75 ng) from each sample was separated in a MOPS/formaldehyde agarose gel and transferred overnight to a Hybond-N$^+$ nylon membrane (Amersham Biosciences). The membrane was hybridized in ExpressHyb hybridization solution (Clontech). The quantification of the transcripts was performed with a Phosphor Imager 425 (Molecular Dynamics). Real-time PCR was performed with an ABI PRISM 7700 instrument (Applied Biosystems). TaqMan probes and primers are given in Supplementary Table 3. PCRs were performed in triplicate using the Universal PCR Master Mix (Applied Biosystems). Messenger RNA was quantified using ten-point calibration curves established by dilution series of the cloned PCR products. Statistical evaluations were done with a two-sided Kruskal–Wallis rank-sum test.

Acknowledgements

We thank C. Charlier and H. Ronne for discussions, M. Laita, B. McTeir, J. Pettersson, A.-C. Svensson and M. Ko¨ping-Ho¨gga”rd for technical assistance, and the Pig Improvement Company for providing DNA samples from Berkshire and Gloucester Old Spot pigs. This work was supported by the Belgian Ministère des Classes Moyennes et de l’Agriculture, the AgriFunGen program at the Swedish University of Agricultural Sciences, the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning, Gentec, the UK Department for Environment, Food and Rural Affairs, the UK Pig Breeders Consortium, and the Biotechnology and Biological Sciences Research Council.

Supplementary information

Table S1: Distribution of genotypes at the Quantitative Trait Nucleotide *IGF2*-intron3-nt3072*G→A* among pig populations strongly selected (+) or not strongly selected (-) for lean growth

		Genotype			
Breed	Lean	*G/G*	*G/A*	*A/A*	Total
European Wild Boar	-	5	0	0	5
European Wild Boar - Uppsala[a]	-	2	0	0	2
Japanese Wild Boar	-	5	0	0	5
Meishan – Roslin[b]	-	11	0	0	11
Berkshire	-	5	0	0	5
Gloucester Old Spot	-	4	0	0	4
Large White – Uppsala[a]	+	0	1	7	8
Large White – Roslin[b]	+	6	1	0	7
Large White – Liège[c]	+	7	0	0	7
Swedish Large White[d]	+	0	0	5	5
Swedish Hampshire[d]	+	0	0	6	6
Swedish Landrace[d]	+	0	0	5	5
Piétrain – Liège[c]	+	0	1	6	7
Duroc	+	0	0	1	1
Total		45	3	30	78

[a]Founder animals in a Wild Boar x Large White intercross (Jeon et al., 1999).

[b]Founder animals in a Large White x Meishan intercross (Walling et al., 1998).

[c]Founder animals in a Piétrain x Large White intercross (Nezer et al., 1999).

[d]Breeding boars that have been tested for QTL segregation in a previous study (Evans et al., 2003). The lack of evidence for QTL segregation strongly suggested that they are all homozygous at the *IGF2* QTN.

Table S2: Primer sequences for long range PCR reactions and sequencing of pig *IGF2*

LTPCR fragments	Names	Sequences
1(3.1kb)	1.MTHE14UP1	5'-CATCCAGCGCCCCTTCTCGGTGAAGTTCGA-3'
	2.MTH_INSUP1	5'-CTTCCCGCAGGATGTAAGCACCAGCCTATCT-3'
	2.MTH_INSUP2	5'-CAGGGTCTGCGTCTGCACGGGCTCA-3'
	3.MTH_INSUP3	5'-GCTGGCAGGAGCCCCACTAGGTCTA-3'
	4.MTH_INSUP4	5'-CCTGCCTTGCACACCCTCTCATAGA-3'
	5.MTH_INSUP5	5'-CTCCCGCCGTTGGAGATGAGAAGCA-3'
	6.INSUP3	5'-CGCTGATGACCCACGGAGATGATCC-3'
	7.INSDN4	5'-GAGCCCACGTCCTCCTGCCGCGAT-3'
	8.MINSUP2	5'-CACCCCCCGCCATGGCCCTGTGGACGCGCCTCCT-3'
	9.INSDN6	5'-AGGGGGGCTTCTCGAGCGGGACCG-3'
	10.MINSDN5	5'-GGCGTAGTTGCAGTAGTTCTCCAGCTGGTAGA-3'
2(4.8kb)	1.MSUP45	5'-GACCGGTGGCTGCTGCGGCTTCCACTCCA-3'
	2.SSCPUP1	5'-CACTGCTGCACAGGGTCACTC-3'
	3.SSCPDN1	5'-CGCCCAGCTCTAGGCCCGG-3'
	4.SSCPUP2	5'-GGAGGCCAGGGCTGGGCAG-3'
	5.SSCPDN2	5'-GGTGGTCACGGGTCTCCCTG-3'
	6.SSCPUP3	5'-GAGTCATGGTGTGACAGCGCC-3'
	7.SSCPDN3	5'-GGGCTGTTGAGCAAAAGGCCAG-3'
	8.SSCPUP4	5'-GACCTGGCTTCGTTTGCCCTG-3'
	9.SSCPDN4	5'-CGGGCCTGACCCCAGCCTC-3'
	10.SSCPUP5	5'-TCCTCCCTGTCCTCCAGGCC-3'
	11.SSCPDN5	5'-AGCCCTGCCTGGGTCCAGGG-3'
	12.SSCPUP6	5'-CAGGTGGACCTCTCTGCAGGG-3'
	13.SSCPDN6	5'-TTCCCCAGGCCTTGGAGACCC-3'

	14.SSCPUP7	5'-CAGCTGGGCCAGGGTTCGTTC-3'
	15.SSCPDN7	5'-CCGCCAGTCTCCACTTTCGAAG-3'
	16.SSCPUP8	5'-CCCTCCAAGTCTGAGATATGGG-3'
	17.SSCPDN8	5'-GTGGGGTCCCCCAAAACCCAC-3'
	18.SSCPUP9	5'-GGCAGCCCTGCAGCTTGCGG-3'
	19.SSCPDN9	5'-GCCCCGTTCTGGTTGCGCATG-3'
	20.SSCPUP10	5'-CCAAAGAAAGCCTCAAGGAGCAG-3'
	21.SSCPDN10	5'-CCCAGGAAGCCTCCCTGTCGG-3'
	22.MSDN45A	5'-CACAGCCATGGGCAACAGAGCAAGGACTTA-3'
3(4.7kb)	1.MSUP47	5'-CACCCTGGGAGCCGTGTTCATGCAGAGAT-3'
	2.SSCPUP11	5'-TCATGCAGAGATGGCGTCCTCC-3'
	3.SSCPDN11	5'-AGGCTAAGTAACATTCCCAAGGCC-3'
	4.SSCPUP12	5'-TTGTGACCCCGGCCATGACATG-3'
	5.SSCPDN12	5'-GGCAACCCCACCCCATGCCC-3'
	6.SSCPUP13	5'-GCCCCAGGAAAGAACGTTCTGG-3'
	7.SSCPDN13	5'-GACGGGTAGAGGACTGGGCAG-3'
	8.SSCPUP14	5'-CTCTGGGGACCGCTGGTGAG-3'
	9.SSCPDN14	5'-TCCTGACCCCAGACCCGCAG-3'
	10.SSCPUP15	5'-CCTGGGCTCGGGAGCCCCC-3'
	11.SSCPDN15	5'-AATTACAGCACACAGCCCAGAGG-3'
	12.SSCPUP16	5'-CCTAAGAGGCAGTGGGGCAATG-3'
	13.SSCPDN16	5'-CCTTGAAACGCAGAGCATTTAAGC-3'
	14.SSCPUP17	5'-AAATGCTACTTCCGTTTCAGCCCC-3'
	15.SSCPDN17	5'-GGGCTCTGCAGCTTCAAGTGTTC-3'
	16.SSCPUP18	5'-CCCCTCCTTAGTGTTGTTATGGC-3'
	17.SSCPDN18	5'-ACCCCTGCACCCTGGGGTGTC-3'
	18.SSCPUP19	5'-TGGGCCCTTCCCCAGCTGCTG-3'
	19.SSCPDN19	5'-TGTCCATCTTGCTTCCAGAGGCAG-3'
	20.SSCPUP20	5'-TCAATCTACTGAGGCCCTGCCC-3'
	21.SSCPDN20	5'-TGCCCTTCCCCAAGCCCCTC-3'
	22.MSDN47	5'-GGACATGCAGCAAGAGGGAACTTCTCACA-3'
4(6.2kb)	1.MSUP1	5'-CACCGGCCAGGCAGTTTCGGCAGAGAGTCT-3'
	2.MSUP2	5'-GCTTCGACTTTGGAGGGGACAGGAA-3'
	3.MSUP3	5'-GCTTCCCTCTCCCCTACTCCCTACAT-3'
	4.MSUP4	5'-GAAGTGTTGGGGGGAGGAGAAGAGT-3'
	5.MSUP5	5'-GTCGGGGTCCTCAGCGGGGAGCT-3'
	6.MSUP6	5'-CCTTACGAGTCCCTGTGTCATGATT-3'
	7.MSUP7	5'-CCAGACTCTACATTAGATGGTGCAAT-3'
	8.MSUP8	5'-GGCTCGCGGGGTGAGTCCGAATCT-3'
	9.MSUP9	5'-GCGCATGTCCAAGCACCAGCAGAA-3'
	10.MSUP10	5'-GACGGGGACTGTGGAAGGCTGTT-3'
	11.MSUP11	5'-CTGGCAGCCCCTCAACCCACAGCT-3'
	12.MSUP12	5'-GGACCGAGCCAGGGACGAGCCT-3'
	13.MSUP13	5'-GCCCGCTCCTCCCAATTGCAGACACGACTT-3'
	14.MSUP14	5'-GCTTAGCGCCCCTGTTTGGCTCCCCCACA-3'
	15.MSUP15	5'-GCTCTCCAAACTCCCTTGCCTATAAT-3'
	16.MSDN1	5'-CTTGCCAGGATCTGGGGTCCTCT-3'
	17.MSDN2	5'-GCCCCTTCGTCGTCCTTCTCACT-3'
	18.MSDN3	5'-CTCTTCTCCTCCCCCCAACACTT-3'
	19.MSDN4	5'-CCCGCTGAGGACCCCGACTGT-3'
	20.MSDN5	5'-CCGGCTCAACACAGTGGAATCATGA-3'
	21.MSDN6	5'-CCAGAAGTCAGTCGGAAGTATGTAT-3'
	22.MSDN7	5'-CCACAAACCCTATTCCAACAACACAA-3'
	23.MSDN8	5'-GGCATTGGTGCCAGTGGCCCTTT-3'
	24.MSDN9	5'-CACAGTCCCCGTCCTTGCACTCA-3'
	25.MSDN10	5'-GAACTGGCCAGCCCCTCGGCACT-3'
	26.MSDN11	5'-CGCTCCCGCGCTGTGAGCCTA-3'
	27.MSDN12	5'-GTGATCTGGAGAAATGGTGAATGTATCT-3'
	28.MSDN13	5'-CGCTAAGCTTGGGTGCCCCTGGCGTCCAAAATTAT-3'
	29.MSDN14	5'-CGCAGCTCCAATAATGGCTCTGTGT-3'
	30.MSDN15	5'-GGCGTTGGCCTGGGATTTGGGAATCAGTTCTGAAT-3'
5(5.5kb)	1.MSUP13	5'-GCCCGCTCCTCCCAATTGCAGACACGACTT-3'
	2.MSUP16	5'-GAATCCGGGCTTCCTAAAATTCAGAA-3'
	3.MSUP17	5'-CTCAGTTCTTCCCGAGGACTTCTTCA-3'
	4.MSUP18	5'-CTGGGGCCCTATCTTACTAGGGTT-3'
	5.MSUP19	5'-GGGTTCTGCCAAGTTTCCCGCTGA-3'
	6.MSUP20	5'-GTCTCTTGCGTTCCCCAAATACACT-3'
	7.MSUP21	5'-GAGCTCCCCGTCGCCTCGCGT-3'
	8.MSUP22	5'-GAGCGCGGGCGAAGTGATTGAT-3'
	9.MSUP23	5'-GCAGGTAGGCTTGGAGCGAGGTT-3'
	10.MSUP24	5'-CCCCATACACTTTCGTACAGCGATT-3'
	11.MSUP25	5'-CGTCCGTCGCTACGCTGCTGACT-3'
	12.MSUP26	5'-GGGCTCCGTCGGCCCAAACCGA-3'
	13.MSUP27	5'-CCCTCTCCGCTTCGCCGTCCCAAAGTGGATTAA-3'
	14.MSDN16	5'-CGTTCTTGAAGAAGTCCTCGGGAAGA-3'
	15.MSDN17	5'-GGTTTCAATTAGTTTATTTTTGGCAAACA-3'
	16.MSDN18	5'-GGGGATGACCGAGGAAGTCTCA-3'
	17.MSDN19	5'-GGGGCAATGCCCAGTCGTTTTCCT-3'
	18.MSDN20	5'-GGGACGGCGGCGAGTGGGTTT-3'
	19.MSDN21	5'-CCTCGCTCCGCCATCAATCACTT-3'
	20.MSDN22	5'-CGGGAACCGGGGAGTCGTCT-3'
	21.MSDN23	5'-CCCGAATCGCTGTACGAAAGTGTAT-3'
	22.MSDN24	5'-CAGCGTAGCGACGGACGGTCAGT-3'
	23.MSDN25	5'-GAAGCGAGGGAGAGGGCGCAACGT-3'
	24.MSDN26	5'-GGGCAGAGATAGTGAGAGACACAGAGTGAACGTGAA-3'
	25.MSDN27	5'-CCCGCCTCCCTGCTGCGTATCGCAAACCGAACA-3'

6(3.6kb)	1.MSUP2728	5'-CCCTCTATCCTTGATACAACAGCTGACCTCACTT-3'
	2.MSUP28	5'-CGGCCCGTCCTCCCCAAACAATCAGACGAGT-3'
	3.MSUP29	5'-GTGAGCTGCTGCGAGGCGACTT-3'
	4.MSUP30	5'-CATACAAGGAGGTGGAAAGCAGT-3'
	5.MSUP31	5'-CCAGCCTGCAAACTGGACATTAGCT-3'
	6.MSUP32	5'-CCCGAATCACTGGCTGACCACA-3'
	7.MSUP33	5'-CACAGGGGGCTTCCGCCTTTGA-3'
	8.MSUP34	5'-CCAAGCTTCGCTCAGCCATAGA-3'
	9.MSUP35	5'-GGGACAGCGACCCCCATGTGAA-3'
	10.MSUP36	5'-CGCTGCTCGTGCTGCTCGTCTT-3'
	11.MSDN28	5'-GCTTTGCGCCCTCCTCATTACACA-3'
	12.MSDN29	5'-GGACTGCTTTCCACCTCCTTGTAT-3'
	13.MSDN30	5'-GGGAGCCTGCCTCAGAGGAGAA-3'
	14.MSDN31	5'-CCTTCGGAGGGCCACCTTCCTCA-3'
	15'.MSDN32	5'-GCAAGCCACCCCTGTGCTGCTCAAA-3'
	16.MSDN33	5'-GGTCCTGCTGGACTAGTCTCTCCT-3'
	17.MSDN34	5'-GACCGGGCCGGGAACTTCACAT-3'
	18.MSDN35	5'-CGGTCCCCGCAGACAAACTGGA-3'
	19.MSDN36	5'-GCACGCCTGCCACTCACGGTCT-3'
	20.MSDNL36	5'-CCCCTCGCACGCCTGCCACTCACGGTCT-3'
7(3.7kb)	1.MSUPL235	5'-CTCCGACTGGAGCAGGTCCTCATCCCTTAGA-3'
	2.MSUP37	5'-CGTTCCTGCAGCGTGACTCGAA-3'
	3.MSUP38	5'-CCTGCTGGCCTCAGAGTGCTTT-3'
	4.MSUP39	5'-GCCTCTTCTGCAGCCTCTACTTA-3'
	5.MSUP40	5'-CTTGCAGCCGAGTCGGAGCTGT-3'
	6.MSUP41	5'-CCACTCTCACTTCTTCTGCTCGA-3'
	7.MSUP42	5'-GTCGACCCCTCCGACCGTGCTT-3'
	8.MSUP43	5'-GTCGCACGCTCGCCAAGGAGCT-3'
	9.MSUP44	5'-CCAATTATCCCCAAGTTACATACCAA-3'
	10.MSDN37	5'-GATGAGGACGGCAAAGCACTCTGA-3'
	11.MSDN38	5'-CCATTTCAGAGGGTGGGTAAGTAGA-3'
	12.MSDN39	5'-GGAAGACGGGAGGAGAGGAGTGA-3'
	13.MSDN40	5'-GCCTCGAGCAGAAGAAGTGAGAGT-3'
	14.MSDN41	5'-CTGCCGAGAGAGGGGCTGCCTTA-3'
	15.MSDN42	5'-GGAAGGGGCCGAGGAGAGACTAGCCTGACAT-3'
	16.MSDN43	5'-GGTCCCTTCCTTTCTCTTGGATGAT-3'
	17.MSDN44	5'-GTGTGGGTACAGGTGCTTTTTAAGTGGAGACTGA-3'
Gaps	1.SSUP7A	5'-GTCAGGGGTCTGGGAGCTGTGGA-3'
	2.SSDN7A	5'-GCCGAGTGGGCGGGCTAGCTCA-3'
	3.MSDN24A1	5'-CCGGGCGCCAACGCCCGGCTCTTATA-3'
	4.GAP1UP1	5'-CAATCCAATCTCGCACCCGCGCACCCCACAGC-3'
	5.GAP1DN1	5'-GCCACCCTGCGGCGCCCGTCAGGCCCCGCT-3'
	6.GAP2UP1	5'-CTAGCGGTTGGGGGTGCAGAAGAGAAGCGAAT-3'
	7.GAP2UP2	5'-GCAGGTAACCGAGGCTGTTGAGCACAGACCT-3'
	8.GAP2UP3	5'-CAGCCCCAGCGAGCCTCCTGTCCAGCTGCA-3'
	9.GAP2UP4	5'-CCTCCTCCTCCTCCAGCCCCAGCGA-3'
	10.GAP2DN1	5'-GTGAGGGGGCTGAGGCCGAGGCTGCAGCTCA-3'

Note: Red-coloured primer sequences were used for both PCR and sequencing reactions. Black-colored primer sequences were used for sequencing reactions

Table S3: Sequence of PCR primers and TaqMan probes used for real-time PCR analysis of pig *IGF2*

Gene	Name	Sequence (5' -> 3')
IGF2	IGF2 forward	CAAGTCCGAGAGGGACGTGT
	IGF2 reverse	CCAGGTGTCATAGCGGAAGAA
	IGF2 probe	CCGACCGTGCTTCCGGACAACTT
GAPDH	*GAPDH* forward	ACCAGGGCTGCTTTTAACTCTG
	GAPDH reverse	TGACAAGCTTCCCGTTCTCC
	GAPDH probe	ACCTCCACTACATGGTCTACATGTTCCAGTATGATT
HPRT	*HPRT* forward	CAGTCAACGGGCGATATAAAAGT
	HPRT reverse	CCAGTGTCAATTATATCTTCAACAATCA
	HPRT probe	ATTGGTGGAGATGATCTCTCAACTTTAACTGGAAA

Figure S1: Neighbor-Joining tree of 18.6 kb of the porcine *IGF2* gene based on 15 sequences classified as representing *q* and *Q* alleles. The analysis was restricted to the region from *IGF2* intron 1 to *SWC9* in the 3'UTR to avoid problems with the presence of recombinant haplotypes. The tree was constructed using MEGA version 2.1 (Kumar et al., 2001) and positions with insertions/deletions were excluded. Bootstrap values (after 1,000 replicates) are reported on the nodes.

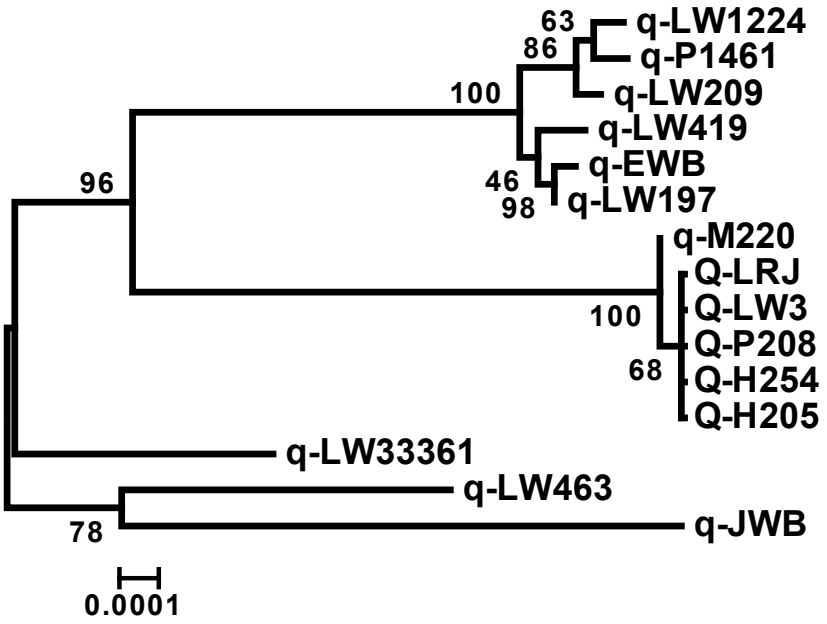
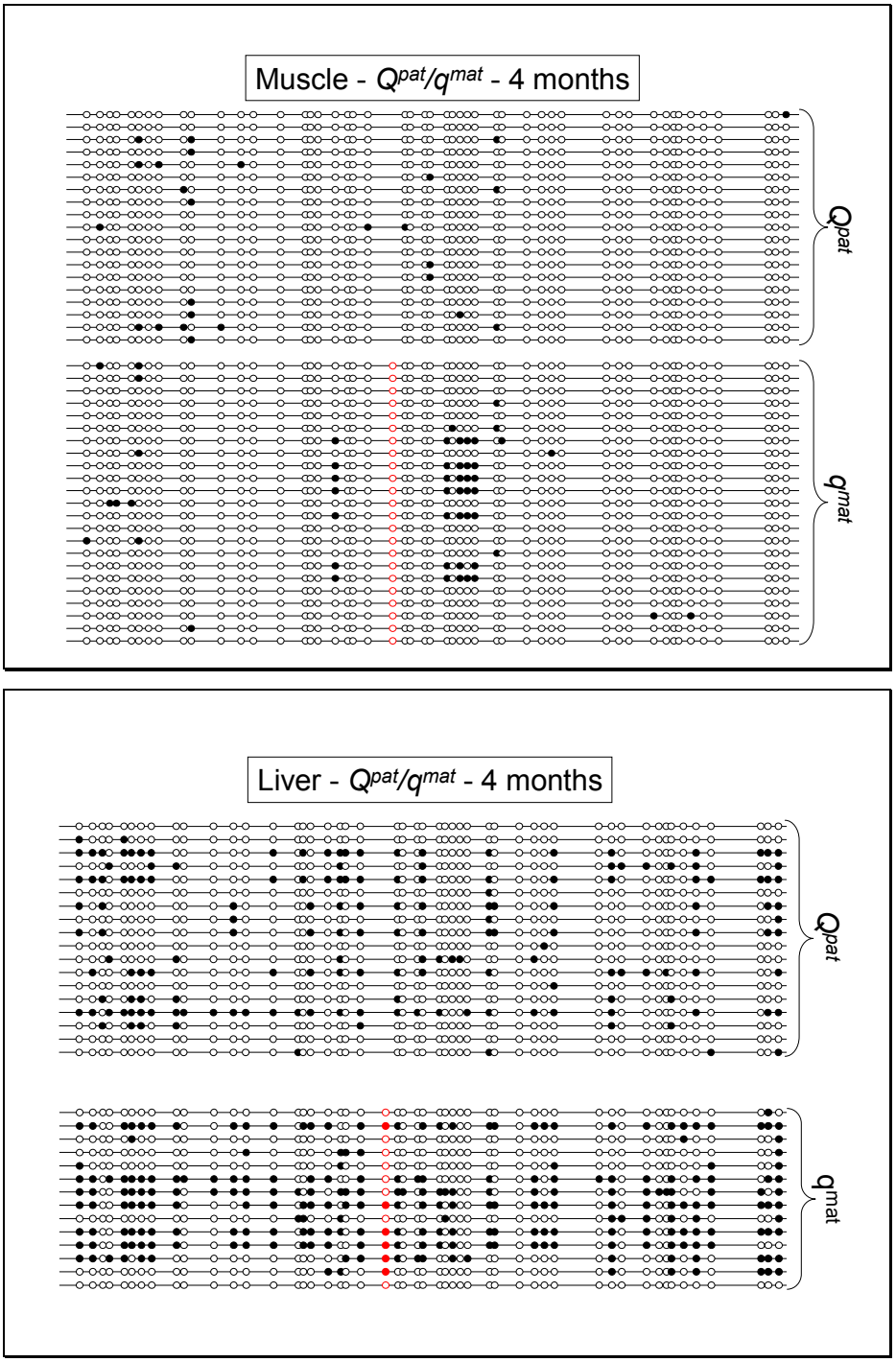
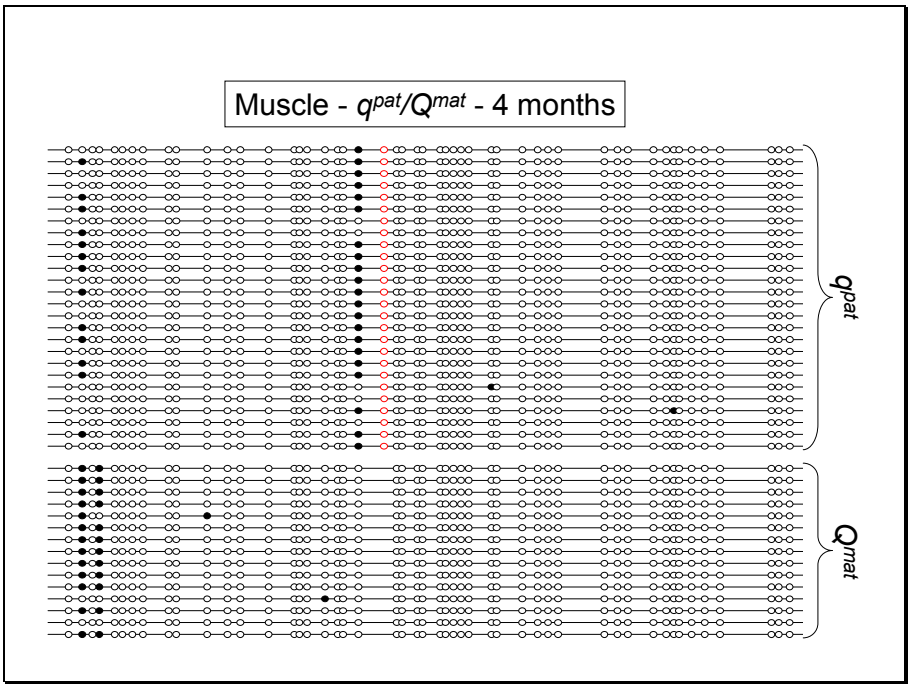


Figure S2: CpG methylation status of individual DNA molecules in liver and skeletal muscle of Q^{pat}/q^{mat} and q^{pat}/Q^{mat} as determined by bisulphite sequencing (see Methods). Empty and filled circles correspond to non-methylated and methylated CpG residues, respectively. Molecules were individualized based on their overall pattern of C to U conversion (see Methods). The position of the CpG including the causative QTN is marked in red.




Muscle - q^{pat}/Q^{mat} - 4 months
q^{pat}
Q^{mat}

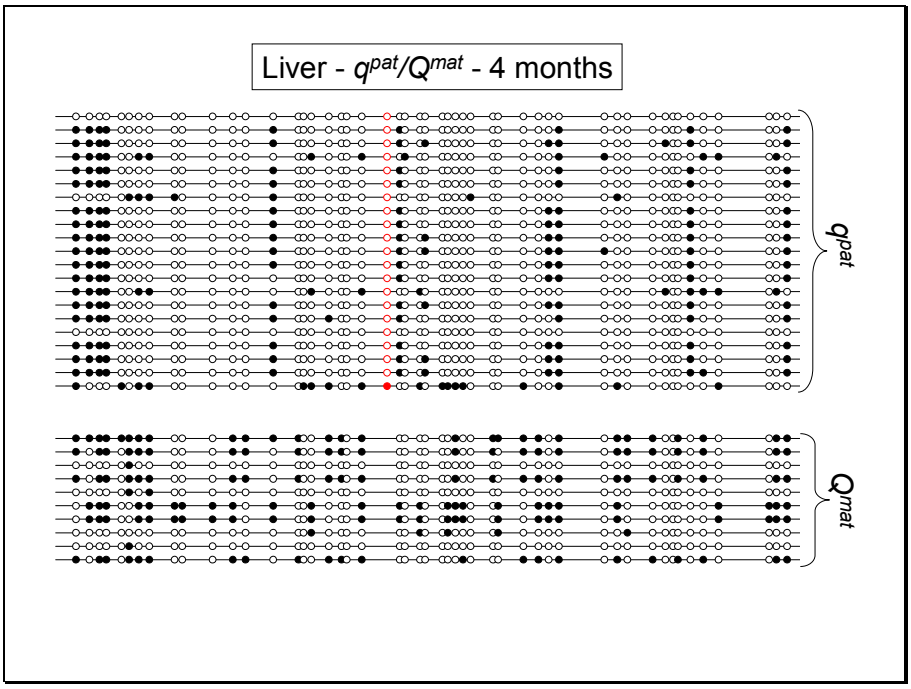

Liver - q^{pat}/Q^{mat} - 4 months
q^{pat}
Q^{mat}

Figure S3: Luciferase assays of reporter constructs using the *TK* promoter. The pig *IGF2* fragments (*Q* and *q*) contained 578 bp from intron 3 (nucleotide 2868 to 3446) including the causative G to A transition at nucleotide 3072. The relative activities compared with the basic TK-LUC reporter are given as means ± SE based on triplicate experiments. An analysis of variance revealed highly significant differences for all pairwise comparisons; ***=P<0.001.

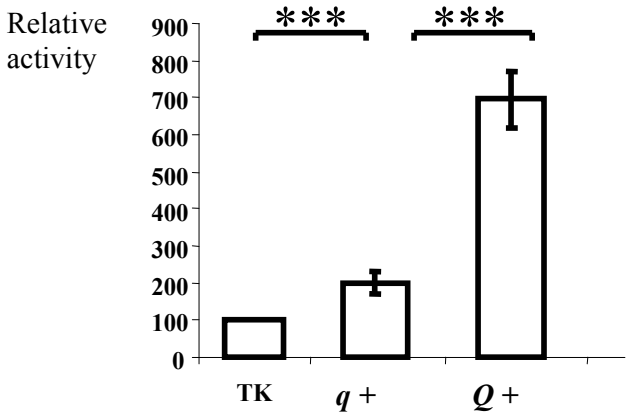


Figure S4: Imprinting analysis of *IGF2* in skeletal muscle (SM) of *qq*, Q^{pat}/q^{mat} and q^{pat}/Q^{mat} animals before birth and at 4 month of age using the highly polymorphic *SWC9* microsatellite (located in *IGF2* 3'UTR) (Glazier et al., 2002). Total RNA was extracted from the *gluteus* muscle using Trizol Reagent (Life Technology), and treated with RNase-free DNase I (Roche Diagnostics GmbH). cDNA was synthesized using the 1st Strand cDNA Synthesis Kit (Roche Diagnostics GmbH). The RT-PCR products were separated and analyzed on an ABI3100 automatic capillary sequencer (Applied Biosystems). The lanes marked "gDNA" correspond to the amplification products obtained from genomic DNA of two individuals heterozygous for the *SWC9* microsatellite (genotype *1/8* or *2/8*). The positions of the corresponding *SWC9* PCR products are marked by black arrows in the electropherograms. The lanes marked "SM cDNA - " correspond to the *SWC9* amplification products obtained by RT-PCR from prenatal and 4 month-old individuals. The QTL genotypes of the corresponding individuals are given in the right column, paternal and maternal alleles being marked in red and green, respectively. In the electropherograms, the paternal *SWC9* alleles are marked by red arrows, the maternal by green arrows. The RT-PCR controls without reverse transcriptase were all negative (data not shown).

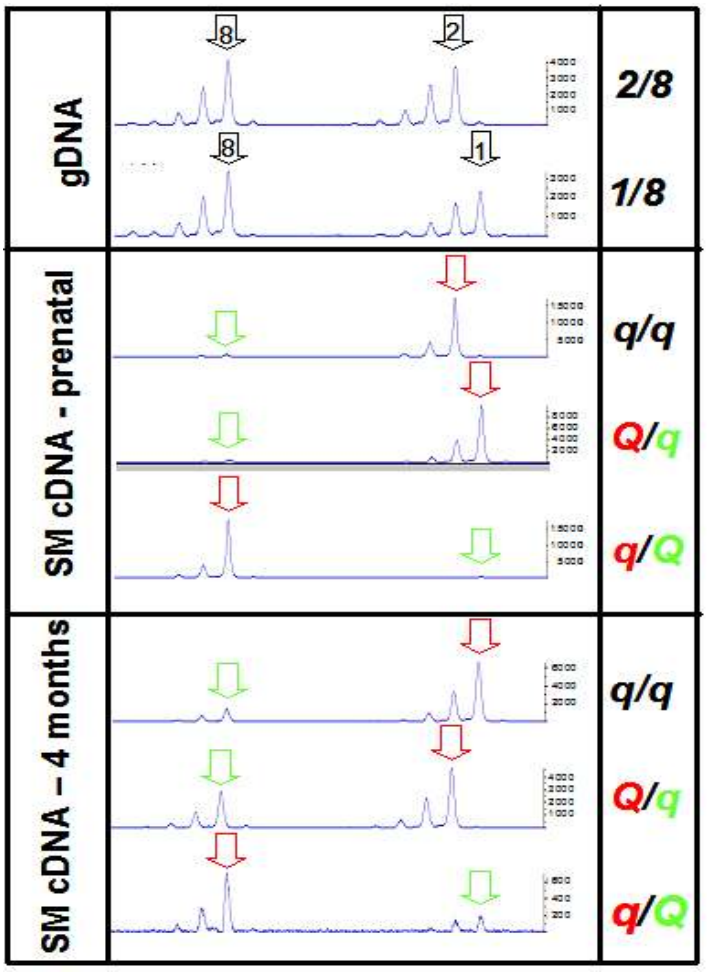
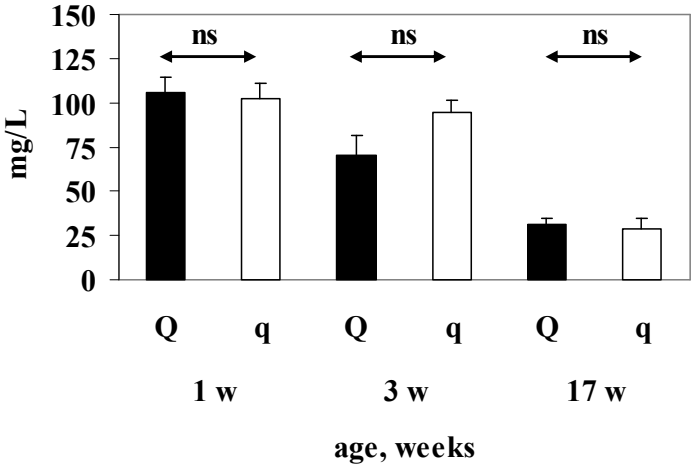


Figure S5: Immunoreactive (ir) IGF2 levels in serum from pigs expressing the *IGF2*q* or *Q* alleles were determined by radioimmunoassay (RIA), using a polyclonal rabbit antibody (Ab) (Olivecrona et al., 1999). IGF-binding proteins were removed before the RIA as previously described (Mohan and Baylink, 1995). Serum was diluted 1:5 in 2.5 M HAc, 0.125 M NaCl, and 100µl of the diluted sample was fractionated over Bio-Gel 10 columns (Bio-Rad Laboratories), using 2M HAc, 0.1M NaCl as eluent. Each serum was separated in duplicate, and duplicate samples of 100 µl from each eluate were lyophilized in the actual assay tube (Minisorp, Nunc), followed by reconstitution in 100 µl 10mM phosphate buffer (pH 7.5) with 1% bovine serum albumin (BSA). RIA was performed in PBS with 1% BSA. IGF1 was added to a final concentration of 25 ng/ml in order to compete for IGF2-affinity to IGFBPs, and this IGF1 concentration does not compete with IGF2 for the Ab. Human recombinant IGF2, a generous gift from Eli Lilly Corp. was used for labeling and as standard.



CHAPTER 4

GENERAL DISCUSSION and PERSPECTIVES

4.1. Power of domestic animal resources for the molecular dissection of complex traits:

The results reported in this thesis vividly illustrate the unique value of domestic animal populations for the identification of genetic variants underlying standing variation for complex phenotypes under selection. Understanding the molecular architecture of complex traits is one of the primary objectives of modern medical and evolutionary genetics. To the best of our knowledge, there is no other system except yeast (Deutschbauer and Davis, 2005), in which regulatory variants contributing to the heritability of quantitative traits have been identified with a comparable degree of confidence. The *IGF2*-Int3-3072G>A mutation is not the only such success in livestock species. Additional regulatory QTN identified using very similar strategies include the *CLPG* and *Texel* mutations, both affecting muscularity by perturbing miRNA-mediated gene regulation in skeletal muscle. Main factors contributing to making livestock populations a unique resource for the positional identification of QTN include the relative genetic isolation and reduced effective population size of breeds limiting the allelic heterogeneity at individual loci including QTL, thus in essence simplifying the problem to solve. In addition, most livestock species are characterized by unusual pedigree structures with common occurrence of very large half-sib pedigrees or harems resulting from the use of artificial reproductive techniques. This allows for accurate genotyping of individual chromosomes at individual QTL by means of marker assisted segregation analysis (MASA). This approach in essence converts a polygenic trait in a series of monogenic entities that are easier to track individually.

It should be noted however that the success cases reported above share the fact that the corresponding QTL effects were at the high end of the distribution. It is becoming increasingly apparent that complex quantitative traits are influenced by what are likely to be hundreds of QTL of which most have very small individual effects on the trait of interest. Dissecting these QTL at the nucleotide level may require prohibitively large samples even in livestock populations and when relying on the tricks applied in this work. How completely it will be possible to understand the molecular architecture of complex traits in the foreseeable future remains to be seen. Continuously improving high throughput sequencing technologies, however, should considerably empower animal as well as other geneticists.

4.2. Asian origin of the *IGF2*-Int3-3072G>A mutation:

Sequencing of 15 distinct *IGF2* haplotypes revealed 258 polymorphisms in a chromosome segment spanning 28.6 Kb, i.e. one polymorphism per 110 base pairs. A phylogenetic tree, build on the basis of the corresponding variant sites, suggested five major branches (Fig. 1).

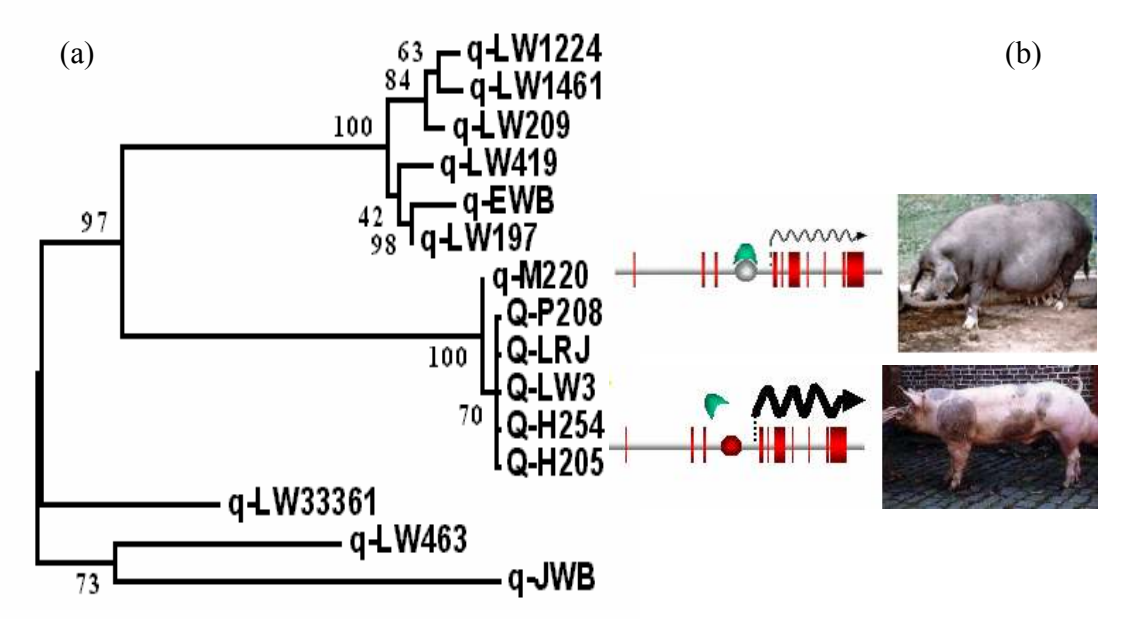


Fig. 1: (a) Neighbor-Joining tree of 18.6 kb of the porcine *IGF2* gene based on 15 sequences classified as representing *q* and *Q* alleles. The analysis was restricted to the region from *IGF2* intron 1 to *SWC9* in the 3'UTR to avoid problems with the presence of recombinant haplotypes. The tree was constructed using MEGA version 2.1 (Kumar et al., 2001) and positions with insertions/deletions were excluded. Bootstrap values (after 1,000 replicates) are reported on the nodes. (b) Effect and modus operandi of the *IGF2*-Int3-3072G>A mutation.

Three of these correspond to haplotype singletons: two q chromosomes originating from Large White and one ungenotyped (QTL) haplotype originating from Japanese Wild Boar. The two remaining branches correspond to two clusters comprising six haplotypes each. The first cluster corresponds the five sequenced Q chromosomes originating respectively from Piétrain, Large White, Landrace or Hampshire (2x), as well as one q chromosome originating from Chinese Meishan. The second cluster corresponds to six q chromosomes originating from Large White and European Wild Boar. The average nucleotide diversity between chromosomes from the two clusters is of the order of 1/120. This is much higher than the typical within species nucleotide diversity in mammals and strongly suggests that the divergence between the corresponding haplotype branches predates domestication, which for pigs is dated at ~9,000 YBP. This finding thus supports the occurrence of multiple independent episodes of domestication of suide subspecies (Giuffra et al., 2000). The fact that the cluster of q chromosomes includes European Wild Boar indicates that the corresponding haplotype was more than likely sampled by domestication of European Wild Boar. The fact that the cluster of Q chromosomes includes a q chromosome of Chinese origin suggests that the mutation that generated the Q allele occurred on a chromosome of Asian ancestry. Whether this event occurred prior or after domestication, before or after importation of

Chinese stock in Europe, remains formally unknown. The recent finding of the *IGF2*-Int3-3072G>A mutation in several Chinese pig breeds (Yang et al., 2006), however, suggests that the mutation predated the wave of migration of Chinese pigs into Europe in the eighteen and nineteen centuries.

4.3. Effect and modus operandi of the *IGF2*-Int3-3072G>A *IGF2* mutation:

The effect of the causative *IGF2* mutation highlights at least two important facets of *IGF2* physiology. The first is that it affects post-natal growth, while *IGF2* is generally assumed to control foetal growth. The fact that the mutation only affects growth after birth, thus without increasing the risk of dystocia, may very well have contributed to its facile dissemination in the pig population (see hereafter). The second is that the causative mutation does not seem to affect *IGF2* expression levels in liver, the major source of circulating IGF2. Accordingly, no effect of QTL genotype on circulating IGF2 levels was observed. The effect of the QTN seems confined to increasing *IGF2* expression levels in skeletal muscle, influencing muscle growth and fat content in a paracrine mode.

Functional analyses of the causative *IGF2* mutation indicate that it may operate by abrogating binding of a trans-acting repressor. EMSA experiments demonstrated formation of a wild-type specific complex using nuclear extracts from murine myoblasts, human embryonic kidney cells and hepatocytes. This suggests that the factor is evolutionary conserved and widely expressed. The molecular basis of the exclusive phenotypic manifestation of the QTL effect in post-natal striated muscle (both skeletal and cardiac) remains to be established. Recently, purification of SILAC-labeled nuclear proteins using biotinilated oligonucleotides followed by mass spectrometric analysis, allowed identification of the corresponding regulatory factor which, reportedly, defines a novel family of trans-acting regulators (Leif Andersson, personal communication).

4.4. On the detection of imprinted QTL in line-crosses:

The detection of the first imprinted QTL by Nezer et al. (1999) and Jeon et al. (1999) spurred many efforts to identify additional parent-of-origin dependent QTL effects in livestock species. Indeed, as alternate microsatellite alleles are generally not fixed in lines of domestic animals (contrary to inbred strains of mice), the two classes of heterozygote genotypes (“12” and “21”) can frequently be distinguished in the intercross generation, allowing testing of the imprinting hypothesis. Early attempts indeed resulted in the identification of a flurry of imprinted QTL in livestock species, suggesting that the contribution of parental imprinting to

the genetic variation for quantitative traits might be more important than initially suspected (de Koning et al., 2000; 2002). Imprinted QTL were even detected in bird species in which imprinted genes have never been observed (Tuiskula et al., 2004). Reporting imprinted QTL has become standard in animal genetics.

It was soon realized however that non fixation of QTL alleles in the intercrossed lines might yield artefactual imprinting effects if not properly accounted for (de Koning et al., 2000; 2002). Moreover, it was recently demonstrated that linkage disequilibrium between QTL and marker loci, which is the typical situation in livestock species, would exacerbate the issue, conferring artefactual imprinting effects to a very significant proportion of Mendelian QTL detected using the basic line-cross model (Sandor and Georges, 2008).

Testing imprinting reduces to measuring the contrast between "12" and "21". The genotypic distribution of the respective dams (transmitting the "1" and "2" alleles respectively) will most of the time differ. As pointed out by Hager et al. (2008), a significant "12" vs "21" contrast, and hence pseudo-imprinting, may therefore also result from maternal effects.

In summary, most imprinted QTL reported to date in livestock species could very well be statistical artefacts rather than genuine involvement of imprinted genes.

Despite these pitfalls, parental imprinting makes a contribution to the heritability of quantitative traits. This is demonstrated by the findings reported in this thesis as well as in the case of the callipyge phenotype in sheep. The latter muscular hypertrophy is characterized by polar overdominance, i.e. phenotypic expression restricted to heterozygous animals inheriting the *CLPG* mutation from their sire (Cockett et al., 1996). Moreover, carefully conducted analyses using F3 crosses between inbred strains of mice suggest that QTL effects with parent-of-origin effects reminiscent of parental imprinting and polar overdominance might indeed be relatively important (Cheverud et al., 2008).

4.5. Utility of the *IGF2*-Int3-3072G>A mutation for marker assisted selection in pig breeding:

The causative *IGF2* mutation has a major effect on muscle mass. At a life weight of ~100 Kg, alternate homozygotes indeed differ by as much as 3 Kg meat. At that stage, heterozygotes have ~ 1 Kg more meat than the qq homozygotes. It was not surprising therefore that most of the European pig breeds, which have undergone intense selection for muscle mass, exhibit high frequencies of the Q allele, testifying for the selective sweep that must have occurred at that locus (Van Laere et al., 2003). This sweep is likely to be relatively recent. Indeed, the French Large-White population that was used to generate the "Sart Tilman" Large-White x

Piétrain inter-cross (the foundation of the research described in this thesis) in the early 1980-ies contributed only q alleles. On the contrary, the Swedish Large White population that was used to generate the Large White x European Wild Boar at approximately the same time contributed mainly Q alleles.

Having identified the causative *IGF2* mutation nevertheless opens ample opportunities for MAS in the boar lines in breeds in which the Q allele is not yet fixed. This remains the situation for a limited number of European breeds, but may be the rule for most of the endogenous breeds in those parts of the world where selection programs are only beginning to be systematically implemented. This is particularly the case in South-East Asia.

The imprinting status of the *IGF2* gene offers unique opportunities for MAS. As only the paternal *IGF2* allele is expressed, the QTL genotype of the maternal allele is to a large extent irrelevant with regards to muscle mass and fat deposition. Thus, selection for the Q allele in the maternal lines is not useful, at least not to improve the carcass composition of the terminal products. Recent evidence, however, indicates that the Q allele has a negative effect on mothering ability (Nadine Buys, personal communication). This suggests that it might be most advantageous to select for the q allele in the dam lines in order to improve their mothering abilities. Thanks to the imprinting status of the QTL, this could be achieved without compromising the carcass quality of the terminal products. Assuming that the effect of the QTL on mothering abilities will be confirmed, this would likely dramatically increase the utility of the *IGF2*-Int3-3072G>A genotyping, especially in European pig populations, in which (as previously mentioned) the Q allele is at high frequency including in maternal lines.

The previous discussion assumes that the maternal allele is totally silent. It is known, however, that imprinting of the *IGF2* gene undergoes some degree of relaxation after birth. Accordingly, we observed some degree of expression of the maternal allele after birth, including in skeletal muscle (Van Laere et al., 2003). Assuming that the causative *IGF2* mutation affects the level of expression of the maternal allele as well, $Q^{Mat}Q^{Pat}$ animals might be slightly more heavily muscled than $q^{Mat}Q^{Pat}$ animals. Indeed, Jeon et al. (1999) provided evidence supporting this assertion. However, the same tendency was not observed in the companion Nezer et al. (1999) study. Further studies are thus needed to clarify this issue. It is likely, however, that the modest effect of the maternal Q allele on carcass merit would not outweigh its negative effect on mothering abilities. Selection of the q allele in dam lines and of the Q allele in sire lines would thus seem as the optimal MAS scenario, exploiting the full potential of this imprinted QTL.

4.6. Recent advances in positional cloning:

The results reported in this thesis are more than five years old. Since then, advances in the field of genomics have made remarkable strides forward. Recent developments would considerably shorten the time required to complete the work presented in this thesis if it had to be repeated.

First of all, a reference genome sequence has or will soon become available for most livestock species. Species with completed genomic sequence include poultry (Hillier et al., 2004) and bovine (Elsik et al. 2009). However, the sequence of pig and sheep are well advanced and, although no definitive draft sequence is available per se, substantial proportions of the pig genome are now accessible in public databases. Unfinished genomes typically come with some level of annotation which can quite easily be augmented from genomic regions of interest especially on the basis of comparative sequence analysis. Thus, the work presented in publication 1 (Amarger et al., 2002) would in fact be readily available for most of the genome.

Generating reference sequences is typically accompanied by shallow sequencing of individuals from multiple breeds in order to identify large number of SNPs and other polymorphisms (Bovine HapMap, Gibbs et al. 2009). These can then be used to generate high density SNP arrays, encompassing for instance 50,000 SNPs. Such a porcine SNP array (PorcineSNP60 Whole-Genome Genotyping Kits) is now available from Illumina Inc. (http://www.illumina.com/). The consequence of that is that the mapping and fine-mapping stages are now typically merged in a single, genome-wide, combined linkage plus LD analysis that results in the direct fine-mapping of the QTL. Thus the studies reported in Nezer et al. 1999, 2002 and 2003 would probably be reported as a single publication and might encompass the fine-mapping of multiple QTL in the genome.

Going from fine-mapping to the identification of the causal gene and mutations would also have been facilitated by recent advances. Analysis of the transcriptome using either array-based approaches or – more recently – direct sequencing allows for the search of "expression QTL" (eQTL). If the eQTL approach had been applied on skeletal muscle of our Piétrain x Large White F2 pedigree we might immediately have noticed the existence of a cis-eQTL on *IGF2*. Moreover, we might have demonstrated a correlation between muscle mass and expression levels of *IGF2* in post-natal skeletal muscle, while the application of conditional correlation measures would have indicated that *IGF2* expression levels and muscle mass might be causally related. This approach, combining QTL and eQTL mapping, is referred to as "genetical genomics" (Jansen, 2003; Georges, 2007).

Finally recent possibilities for massive parallel resequencing (Mardis 2008) would have greatly accelerated the resequencing of specific “progeny-tested” haplotypes. Regions of interest can either be selected by long range PCR or be captured by hybridization performed either in solution or on solid support and sequenced on ultra high throughput sequencing instruments. If necessary, sequences originating from distinct individuals can be identified by means of “multiplex identifier” tags.

References

Ainscough J. F-X., Rosalind M. J., Sheila C. B. and Azim Surani M. A skeletal muscle-specific mouse Igf2 repressor lies 40 kb downstream of the Gene. *Development* **127**: 3923.

Amarger V., Gauguier D., Yerle M., Apiou F., Pinton P., Giraudeau F., Monfouilloux S., Lathrop M., Dutrillaux B., Buard J. and Vergnaud G. 1998. Analysis of distribution in the human, pig, and rat genomes points toward a general subtelomeric origin of minisatellite structures. *Genomics* **52**: 62.

Amarger V., Nguyen M., Van Laere A.-S., Braunschweig M., Nezer C., Georges M. and Andersson L. 2002. Comparative sequence analysis of the *Insulin-IGF2-H19* gene cluster in pigs. *Mamm. Genome* **13**: 388.

Andersson L., Archibald A., Ashburner M., Audun S., Barendse W., Bitgood J. et al. 1996. Comparative genome organization of vertebrates. The First International Workshop on Comparative Genome Organization. *Mamm. Genome* **7**: 717.

Andersson L. 2001. Genetic dissection of phenotypic diversity in farm animals. *Nat. Rev. Genet*. **2**: 130.

Andersson L. 2009. Genome-wide association analysis in domestic animals: a powerful approach for genetic dissection of trait loci. *Genetica* **136**: 341.

Anderson S. I., Lopez-Corrales N. L., Gorick B. and Archibald A. L. 2000. A large fragment porcine genomic library resource in a BAC vector. *Mamm. Genome* **11**: 811.

Andrews N. C. and Faller D. V. 1991. A rapid micropreparation technique for extraction of DNA-binding proteins from limiting numbers of mammalian cells. *Nucleic Acids Res*. **19**: 2499.

Baker J., Liu J. P., Robertson E. J. and Efstratiadis A. 1993. Role of insulin-like growth factors in embryonic and postnatal growth. *Cell* **75**: 73.

Bell A. C. and Felsenfeld G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the *IGF2* gene. *Nature* **405**: 482.

Bennett S. T., Lucassen A. M., Goug S. C., Powell E. E., Undlien D. E., Pritchard L. E., Merriman M. E., Kawaguchi Y., Dronsfield M. J., Pociot F. et al. 1995. Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the *insulin* gene minisatellite locus. *Nat. Genet*. **9**: 284.

Bestor T. H. and Ingram V. M. 1983. Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proc Natl Acad Sci USA* **80**: 5559.

Bestor T. H. 1988. Cloning of a mammalian DNA methyltransferase. *Gene* **74**: 9.

Brown K. W., Villar A. J., Bickmore W., Clayton-Smith J., Catchpoole D., Maher E. R. and Reik W. 1996. Imprinting mutation in the Beckwith-Wiedemann syndrome leads to biallelic *IGF2* expression through an *H19*-independent pathway. *Hum. Mol. Genet*. **5**: 2027.

Cheverud J. M., Hager R., Roseman C., Fawcett G., Wang B. and Wolf J. B. 2008. Genomic imprinting effects on adult body composition in mice. *PNAS* **105**: 4253.

Clop A., Marcq F., Takeda H., Pirottin D., Tordoir X., Bibé B., Bouix J., Caiment F., Elsen J.M., Eychenne F., Larzul C., Laville E., Meish F., Milenkovic D., Tobin J., Charlier C. and Georges M. 2006. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat. Genet.* **38**: 813.

Cockett N. E., Jackson S. P., Shay T. L., Farnir F., Berghmans S., Snoder G. D., Nielsen D. M. and Georges M. 1996. Polar overdominance at the ovine *callipyge* locus. *Science* **273**: 236.

Cohen-Zinder M., Seroussi E., Larkin D. M., Loor J. J., Everts-vander Wind A., et al. 2005. Identification of a missense mutation in the bovine *ABCG2* gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res.* **15**: 936.

Constancia M., Dean W., Lopes S., Moore T., Kelsey G. and Reik W. 2000. Deletion of a silencer element in *IGF2* results in loss of imprinting independent of *H19*. *Nat. Genet.* **26**: 203.

Cui H., Niemitz E. L., Ravenel J. D., Onyango P., Brandenburg S. A., Lobanenkov V. V. and Feinberg A. P. 2001. Loss of imprinting of Insulin-like Growth factor-II in Wilms'tumor commonly involves altered methylation but not mutations of CTCF or its binding site. *Cancer Res*. **61**: 4947.

Cui H., Cruz-Correa M., Giardiello F. M., Hutcheon D. F., Kafonek D. R., Brandenburg S., Wu Y., He X., Powe N. R. and Feinberg A. P. 2003. Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. *Science* **299**: 1753.

Darvasi A. and Pisante-Shalom A. 2002. Complexities in the genetic dissection of quantitative trait loci. *Trends Genet*. **18**: 489.

De Chiara T. M., Efstratiadis A. and Robertson E.J. 1990. A growth-deficiency phenotype in heterozygous mice carrying an insulin-like growth factor II gene disrupted by targeting. *Nature* **345**: 78.

de Koning D. J., Rattink A. P., Harlizius B., van Arendonk J. A., Brascamp E. W. and Groenen M. A. 2000. Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proc. Natl. Acad. Sci. USA* **97**: 7947.

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

Dekkers J. C. M. and Hospital F. 2002. Multifactorial genetics: The use of molecular genetics in the improvement of agricultural populations. *Nat. Rev. Genet.* **3**: 22.

Delaval K., Wagschal A. and Feil R. 2006. Epigenetic deregulation of imprinting in congenital diseases of aberrant growth. *Bioessays* **28**: 453.

Deutschbauer A. M. and Davis R. W. 2005. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat. Genet.* **37**: 1333.

Doerge R. W. and Churchill G. A. 1996. Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**: 285.

Drewell R. A., Brenton J. D., Ainscough J. F.-X., Barton S. C., Hilton K. J., Arney K. L., Dandolo L. and Surani M. A. 2000. Deletion of a silencer element disrupts *H19* imprinting independently of a DNA methylation epigenetic switch. *Development* **127**: 3419.

Dubchak I., Brudn M., Loots G. G., Mayor C., Pachter L., Rubin E. M. and Frazer K. A. 2000. Active conservation of noncoding sequences revealed by 3-way species comparisons. *Genome Res.* **10**: 130.

Eden S., Constancia M., Hashimshony T., Dean W., Goldstein B., Johnson A. C., Keshet I., Reik W. and Cedar H. 2001. An upstream repressor element plays a role in IGF2 imprinting. *EMBO J*. **20**: 3518.

Eggenschwiler J., Ludwig T., Fisher P., Leighton P. A., Tilghman S. M. and Efstratiadis A. 1997. Mouse mutant embryos overexpressing IGF-II exhibit phenotypic features of the Beckwith-Wiedemann and Simpson-Golabi-Behmel syndromes. *Genes Dev*. **11**: 3128.

Eggermann T., Eggermann K. and Scho¨ nherr N. 2008. Growth retardation versus overgrowth: Silver-Russell syndrome is genetically opposite to Beckwith-Wiedemann syndrome. *Trends Genet*. **24**: 195.

Elsik C. G. et al. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522.

Engemann S., Strödicke M., Paulsen M., Franck O., Reinhardt R., Lane N., Reik W. and Walter J. 2000. Sequence and functional comparison in the Beckwith-Wiedemann region: implications for a novel imprinting center and extended imprinting. *Hum. Mol. Genet*. **9**: 2691.

Engemann S., El-Maarri O., Hajkova P., Oswald J. and Walter J. 2002. in Methods in Molecular Biology Vol. 181 (ed. Ward, A.) (Humana Press, Totowa, New Jersey).

Evans G. J. et al. 2003. Identification of quantitative trait loci for production traits in commercial pig populations. *Genetics* **164**: 621.

Ewing B., Hillier L., Wendl M. and Green P. 1998. Basecalling of automated sequencer traces using phred. I.Accuracy assessment. *Genome Res*. **8**: 175.

Farnir F., Grisart B., Coppieters W., Riquet J., Berzi P., Cambisano N., Karim L., Mni M., Moisio S., Simon P., Wagenaar D., Vilkki J. and Georges M. 2002. Simultaneous mining of linkage and linkage disequilibrium to fine-map QTL in outbred half-sib pedigrees: revisiting the location of a QTL with major effect on milk production on bovine chromosome 14. *Genetics* **161**: 275.

Feinberg A. P. and Tycko B. 2004. The history of cancer epigenetics. *Nat Rev Cancer* **4**: 143.

Feinberg A. P., Ohlsson R. and Henikoff F. 2006. The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.* **7**: 21.

Ferguson-Smith A. C. 2000. Genetic imprinting: Silencing elements have their say. *Current Biology* **10**: 872.

Florini J. R., Ewton D. Z. and McWade F. J. 1995. IGFs, muscle growth, and myogenesis. *Diabetes Rev*. **3**: 73.

Florini J. R., Ewton D. Z. and Coolican S. A. 1996. Growth hormone and the insulin-like growth factor system in myogenesis. *Endocr Rev*. **17**: 481.

Freking B. A. et al. 2002. Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. *Genome Res*. **12**: 1496.

Frevel M. A., Sowerby S. J., Petersen G. B. and Reeve A. E. 1999. Methylation sequencing analysis refines the region of *H19* epimutation in Wilms tumor. *J. Biol. Chem*. **274**: 29331.

Fujii J., Otsu K., Zorzato F., de Leon S., Khanna V. K., Weiler J. E., O'Brien P. J. and MacLennan D. H. 1991. Identification of a mutation in porcine *ryanodine* receptor associated with malignant hyperthermia. *Science* **253**: 448.

Galloway S. M. et al. 2000. Mutations in an oocyte-derived growth factor gene (*BMP15*) cause increased ovulation rate and infertility in a dosage-sensitive manner. *Nat. Genet.* **25**: 279.

Gardiner-Garden M. and Frommer M. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol*. **196**: 261

Georges M. 2007. Mapping, fine mapping, and molecular dissection of quantitative trait Loci in domestic animals. *Annu Rev Genomics Hum Genet*. **8**: 131.

Giannoukakis N., Deal C., Paquette J., Goodyer C. G. and Polychronakos C. 1993. Parental genomic imprinting of the human *IGF2* gene. *Nat. Genet*. **4**: 98.

Gibbs R. A. et al. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**: 528.

Giuffra E. J., Kijas M. H., Amarger V., Carlborg O., Jeon J.-T. and Andersson L. 2000. The origin of the Domestic Pig: Independent Domestication and Subsequent Introgression. *Genetics* **154**: 1785.

Glazier A. M., Nadeau J. H. and Aitman T. J. 2002. Finding genes that underlie complex traits. *Science* **298**: 2345.

Goddard M. E. and Hayes B. J. 2007. Genomic selection. *J. Anim. Breed. Genet*. **124**: 323

Gordon D., Abajian C. and Green P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res*. **8**: 195.

Greally J. M., Guinness M. E., McGrath J. and Zemel S. 1997. Matrix-attachment regions in the mouse chromosome 7F imprinted domain. *Mamm. Genome* **8** : 805.

Grisart B., Coppieters W., Farnir F., Karim L., Ford C., et al. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. *Genome Res.* **12**: 222.

Grisart B., Farnir F., Karim L., Cambisano N., Kim J. J., et al. 2004. Genetic and functional confirmation of the causality of the *DGAT1* K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc. Natl. Acad. Sci. USA* **101**: 2398.

Grobet L. et al. 1997. A deletion in the bovine *myostatin* gene causes the double-muscled phenotype in cattle. *Nat. Genet*. **17**: 71.

Gudrun E., Sayeda N. M., Gill Bell A.-A., Wakeling L. E., Kingsnorth A., Stanier P., Jauniaux E. and Bennett S. T. 2001. Evidence That *Insulin* is Imprinted in the Human Yolk Sac. *Diabetes* **50**: 199.

Hager R., Cheverud J. M. and Wolf J. B. 2008. Maternal Effects as the Cause of Parent-of-Origin Effects That Mimic Genomic Imprinting. *Genetics* **178**: 1755.

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

Hatada I., Ohashi H., Fukushima Y., Kaneko Y., Inoue M., Komoto Y., Okada A., Ohishi S., Nabetani A., Morisaki H., Nakayama M., Niikawa N. and Mukai T. 1996. An imprinted gene p57KIP2 is mutated in Beckwith-Wiedemann syndrome. *Nat. Genet*. **14**: 171.

Heude B., Ong K. K., Luben R., Wareham N. J. and Sandhu M. S. 2007. Study of association between common variation in the insulin-like growth factor 2 gene and indices of obesity and body size in middle-aged men and women. *J Clin Endocrinol Metab*. **92**: 2734.

Hillier L. W. et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695.

Holmgren C., Kanduri C., Dell G., Ward A., Mukhopadhya R., Kanduri M., Lobanenkov V. and Ohlsson R. 2001. CpG methylation regulates the *IGF2/H19* insulator. *Current Biol*. **11**: 1128.

Holthuizen P., van der Lee F. M., Ikejiri K., Yamamoto M. and Sussenbach J. S. 1990. Identification and initial characterization of a fourth leader exon and promoter of the human *IGF2* gene. *Biochim. Biophys. Acta*. **1087**: 341.

Ishihara K., Hatano N., Furuumi H., Kato R., Iwaki T., Miura K., Jinno Y. and Sasaki H. 2000. Comparative genomic sequencing identifies novel tissue-specific enhancers and sequence elements for methylation-sensitive factors implicated in *IGF2/H19* imprinting. *Genome Res*. **10**: 664.

Jansen R. C. 2003. Studying complex biological systems using multifactorial perturbation. *Nat. Rev. Genet.* **4**: 145.

Jareborg N., Birney E. and Durbin R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res*. **9**: 815.

Jareborg N. and Durbin R. 2000. Alfresco-a workbench for comparative genomic sequence analysis. *Genome Res*. **10**: 1148.

Jeon J. -T., Carlborg Ö., Törnsten A., Giuffra E., Amarger V., Chardon P., Andersson-Eklund L., Andersson K., Hansson I., Lundström K. and Andersson L. 1999. A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the *IGF2* locus. *Nat. Genet.* **21**: 157.

Jeltsch A., Nellen W. and Lyko F. 2006. Two substrates are better than one: dual specificities for *Dnmt2* methyltransferases. *TRENDS in Biochemical Sciences*. **31**: 306.

Kaffer C. R., Srivastava M., Park K. Y., Ives E., Hsieh S., Batlle J., Grinberg A., Huang S. P. and Pfeifer K. A. 2000. Transcriptional insulator at the imprinted *H19/IGF2* locus. *Genes Dev*. **14**: 1908.

Kashuk C., Sengupta S., Eichler E. and Chakravarti A. 2002. ViewGene: a graphical tool for polymorphism visualization and characterization. *Genome Res*. **12**: 333.

Kim Y. J., Yoon J.-H., Kim C. Y., Kim L. H., Park B. L., Shin H. D. and Lee H. S. 2006. IGF2 polymorphisms are associated with hepatitis B virus clearance and hepatocellular carcinoma. *BBRC* **346**: 38.

King M. C. and Wilson A. C. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107.

Kumar S., Tamura K., Jakobsen I. B. and Nei M. 2001. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**: 1244.

Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860.

LeRoith D. and Bondy C. 1996. Growth hormones and cytokines in health and diseases. ISBN: 0762300914.

LeStunff C., Fallin D. and Bougnères P. 2001. Paternal transmission of the very common class I INS VNTR alleles predisposes to childhood obesity. *Nat. Genet*. **29**: 96.

Liu J-P., Baker J., Perkins A. S., Robertson E. J. and Efstratiadis A. 1993. Mice carrying null mutations of the genes encoding insulin-like growth factor I (Igf-1) and type1 receptor (Igf1r). *Cell* **75**: 59.

Lynch M. and Walsh B. 1998. Genetics and Analysis of Quantitative Traits. ISBN: 0878934812.

Mackay T. F. C. 2001. Quantitative trait loci in Drosophila. *Nat. Rev. Genet*. **2**: 11.

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

Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S. and Dubchak I. 2000. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**: 1046.

Meuwissen T. H. E. and Goddard M. E. 2001. Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol*. **33**: 605.

Milan D. et al. 2000. A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* **288**: 1248.

Mineo R., Fichera E., Liang S.-J. and Fujita-Yamaguchi Y. 2000. Promoter usage for insulin-like growth factor-II in cancerous and benign human breast, prostata, and bladder tissues, and confirmation of a 10th exon. *Biochem. Biophys. Res. Commun*. **268**: 886.

Mohan S. and Baylink D. J. 1995. Development of a simple valid method for the complete removal of insulin-like growth factor (IGF)-binding proteins from IGFs in human serum and

other biological fluids: comparison with acid-ethanol treatment and C18 Sep-Pak separation. *J. Clin. Endocrinol. Metab.* **80**: 637.

Monk D., Sanches R., Arnaud P., Apostolidou S., Hills F. A., Abu-Amero S., Murrell A., Friess H., Reik W., Stanier P., Constaˆncia M. and Moore G. E. 2006. Imprinting of IGF2 P0 transcript and novel alternatively spliced INS-IGF2 isoforms show differences between mouse and human. *Human Molecular Genetics* **15**: 1259.

Mulsant P. et al. 2001. Mutation in bone morphogenetic protein receptor-IB is associated with increased ovulation rate in Booroola Merino ewes. *Proc. Natl Acad. Sci. USA*. **98**: 5104.

Murrell A., Heeson S., Bowden L., Constância M., Dean W., Kelsey G. and Reik W. 2001. An intragenic methylated region in the imprinted Igf2 gene augments transcription. *EMBO Rep* **2**:1101.

Musarò A., McCullagh K., Paul A., Houghton L., Dobrowolny G., Molinaro M., Barton E. R., Sweeney H. L. and Rosenthal N. 2001. Localized Igf-1 transgene expression sustains hypertrophy and regeneration in senescent skeletal muscle. *Nat. Genet*. **27**: 195.

Nakagawa H., Chadwick R. B., Peltomaki P., Plass C., Nakamura Y. and de La Chapelle A. 2001. Loss of imprinting of the insulin-like growth factor II gene occurs by biallelic methylation in a core region of H19-associated CTCF-binding sites in colorectal cancer. *Proc. Natl Acad. Sci U.S.A*. **98**: 591.

Nezer C., Moreau L., Brouwers B., Coppieters W., Detilleux J., Hanset R., Karim L., Kvasz A., Leroy P. and Georges M. 1999. An imprinted QTL with major effect on muscle mass and fat deposition maps to the *IGF2* locus in pigs. *Nat. Genet*. **21**: 155.

Nezer C., Collette C., Moreau L., Brouwers B., Kim J. J., Giuffra E., Buys N., Andersson L. and Georges M. 2003. Haplotype sharing refines the location of an imprinted quantitative trait locus with major effect on muscle mass to a 250-kb chromosome segment containing the porcine IGF2 gene. *Genetics*. **165**: 277.

Nezer C., Moreau L., Wagenaar D. and Georges M. 2002. Results of a whole genome scan targeting QTL for growth and carcass traits in a Piétrain x Large White intercross. *Genet. Sel. Evol.* **34**: 371.

Nickerson D., Tobe V. O. and Taylor S. L. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescent-based resequencing. *Nucleic Acids Res*. **25**: 2745.

Niemitz E. L., DeBaun M. R., Fallon J., Murakami K., Kugoh H., Oshimura M. and Feinberg A. P. 2004. Microdeletion of LIT1 in familial Beckwith-Wiedemann syndrome. *Am J Hum Genet*. **75**: 844.

Ohlsen S. M., Lugenbeel K. A. and Wong E. A. 1994. Characterization of the linked ovine insulin and insulin-like growth factor-II genes. *DNA Cell Biol*. **13**: 377.

Ohlsson R., Renkawitz R. and Lobanenkov V. 2001. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in Genet*. **17**: 520.

Okamura K., Hagiwara-Takeuchi Y., Li T., Vu T. H., Hirai M., Hattori M., Sakaki Y., Hoffman A. R. and Ito T. 2000. Comparative genome analysis of the mouse imprinted gene impact and its nonimprinted human homolog IMPACT: toward the structural basis for species-specific imprinting. *Genome Res*. **10**: 1878.

Olivecrona H. *et al.* 1999. Acute and short-term effects of growth hormone on insulin-like growth factors and their binding proteins: serum levels and hepatic messenger ribonucleic acid responses in humans. *J. Clin. Endocrinol. Metab.* **84**: 553.

Onyango P., Miller W., Lehoczky J., Leung C. T., Birren B., Wheelan S., Dewar K. and Feinberg A. P. 2000. Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. *Genome Res*. **10**: 1697.

Pacher M., Seewald M. J., Mikula M., Oehler S., Mogg M., Vinatzer U., Eger A., Schweifer N., Varecka R., Sommergruber W., Mikulits W. and Schreiber M. 2007. Impact of constitutive IGF1/IGF2 stimulation on the transcriptional program of human breast cancer cells. *Carcinogenesis* **28**: 49.

Pailhoux E. et al. 2001. A 11.7-kb deletion triggers intersexuality and polledness in goats. *Nat. Genet*. **29**: 453.

Paquette J., Giannoukakis N., Polychronakos C., Vafiadis P. and Deal C. 1998. The *INS* 5' variable number of tandem repeats is associated with *IGF2* expression in humans. *J. Biol. Chem*. **273**: 14158.

Paulsen M., El-Maarri O., Engemann S., Strödicke M., Franck O., Davies K., Reinhardt R., Reik W. and Walter J. 2000. Sequence conservation and variability of imprinting in the Beckwith-Wiedemann syndrome gene cluster in human and mouse. *Hum. Mol. Genet*. **9**: 1829.

Perier R. C., Praz V., Junier T., Bonnard C. and Bucher, P. 2000. The eukaryotic promoter database (EPD). *Nucleic Acids Res*. **28**: 302.

Powell-Braxton L., Hollingshead P., Warburton C., Dowd M., Pitts-Meek S., Dalton D., Gillett N. and Stewart T. A. 1993. IGF-I is required for normal embryonic growth in mice. *Genes Dev*. **7**: 2609.

Pugliese A., Zeller M., Fernandez A. Jr., Zalcberg L. J., Bartlett R. J., Ricordi C., Pietropaolo M., Eisenbarth G. S., Bennett S. T. and Patel D. D. 1997. The *insulin* gene is transcribed in

the human thymus and transcription levels correlated with allelic variation at the *INS* VNTR-IDDM2 susceptibility locus for type 1 diabetes. *Nat. Genet*. **15**: 293.

Rahman N. 2005. Mechanisms predisposing to childhood overgrowth and cancer. *Curr Opin Genet* **15**: 227.

Reed M. R., Huang C.-F., Riggs A. D. and Mann J. R. 2001. A Complex Duplication Created by Gene Targeting at the Imprinted *H19* Locus Results in Two Classes of Methylation and Correlated *IGF2* Expression Phenotypes. *Genomics* **74**: 186.

Reik W., Brown K. W., Schneid H., Le Bouc Y., Bickmore W. and Maher E. R. 1995. Imprinting mutations in the Beckwith-Wiedemann syndrome suggested by altered imprinting pattern in the *IGF2-H19* domain. *Hum. Mol. Genet*. **4**: 2379.

Reik W., Murrell A., Lewis A., Mitsuya K., Umlauf D., Dean W., Higgins M. and Feil R. 2004. Chromosome Loops, Insulators, and Histone Methylation: New Insights into Regulation of Imprinting in Clusters. *Cold Spring Harb Symp Quant Biol*. **69**: 29.

Reik W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**: 425.

Rietveld L. E., Holthuizen P. E. and Sussenbach J. S. 1997. Identification of a key regulatory element for the basal activity of the human insulin-like growth factor II gene promoter P3. *Biochem. J*. **327**: 689.

Riquet J., Coppieters W., Cambisano N., Arranz J. J., Berzi P., Davis S. K., Grisart B., Farnir F., Karim L., Mni M., Simon P., Taylor J. F., Vanmanshoven P., Wagenaar D., Womack J. E. and Georges M. 1999. Identity-by-descent fine-mapping of QTL in outbred populations: application to milk production in dairy cattle. *Proc. Natl. Acad. Sci. USA* **96**: 9252.

Rodriguez S., Gaunt T. R. and Day I. N. M. 2007. Molecular genetics of human growth hormone, insulin-like growth factors and their pathways in common disease. *Hum Genet*. **122**: 1.

Rothschild M. F. and Ruvinsky A. 1998. The genetics of the pig. ISBN: 9780851992297.

Rudd M. F., Webb E. L., Matakidou A., Sellick G. S., Williams R. D., Bridle H., Eisen T., Houlston R. S. 2006. Variants in the GH-IGF axis confer susceptibility to lung cancer. *Genome Res.* **16**: 693.

Sandor C. and Georges M. 2008. On the Detection of Imprinted Quantitative Trait Loci in Line Crosses: Effect of Linkage Disequilibrium. *Genetics* **180**: 1167.

Sandovici I., Leppert M., Hawk P. R., Suarez A., Linares Y. and Sapienza C. 2003. Familial aggregation of abnormal methylation of parental alleles at the IGF2/H19 and IGF2R differentially methylated regions. *Human Molecular Genetics* **12**: 1569.

Shibata H., Yoda Y., Kato R., Ueda T., Kamiya M., Hiraiwa N., Yoshiki A., Plass C., Pearsall R. S., Held W. A., Muramatsu M., Sasaki H., Kusakabe M. and Hayashizaki Y. 1998. A methylation imprint mark in the mouse imprinted gene *Grf1/Cdc25Mm* locus shares a common feature with the *U2afbp-rs* gene: an association with a short tandem repeat and a hypermethylated region. *Genomics* **49**: 30.

Sjogren K., Liu J. L., Blad K., Skrtic S., Vidal O., Wallenius V., LeRoith D., Tornell J., Isaksson O. G., Jansson J. O. and Ohlsson C. 1999. Liver-derived insulin-like growth factor I (IGF-I) is the principal source of IGF-I in blood but is not required for postnatal body growth in mice. *Proc Natl Acad Sci USA* **96**: 7088.

Smits G., Andrew J Mungall A. J., Sam Griffiths-Jones S., Smith P., BeuryD., Matthews L.,Rogers J., Pask A. J., Shaw G., VandeBerg J. L., McCarrey J. R., the SAVOIR Consortium, Renfree M. B., Reik W. and Dunham I. 2008. Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. *Nat. Genet.* **40**: 971.

Sparago A., Cerrato F., Vernucci M., Ferrero G. B., Silengo M. C. and Riccio A. 2004. Microdeletions in the human H19 DMR result in loss of IGF2 imprinting and Beckwith-Wiedemann syndrome. *Nat. Genet*. **36**: 958.

Steele M. R. and Georges M. 1991. Generation of bovine multisite haplotypes using random cosmid clones. *Genomics* **10**: 889.

Terwilliger J. D. and Weiss K. M. 1998. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.* **9**:578.

Thorvaldsen J. L., Duran K. L., Bartolomei M. S. 1998. Deletion of the *H19* differentially methylated domain results in loss of imprinted expression of *H19* and *IGF2*. *Genes Dev*. **12**: 3693.

Tilghman S. M., Bartolomei M. S., Webber A. L., Brunkow M. E., Saam J., Leighton P. A., Pfeifer K. and Zemel S. 1993. Parental imprinting of the *H19* and *IGF2* genes in the mouse. *Cold Spring Harb Symp Quant Biol* **58**: 287.

Tremblay K. D., Saam J. R., Ingram R. S., Tilghman S. M. and Bartolomei M. S. 1995. A paternal-specific methylation imprint marks the alleles of the mouse *H19* gene. *Nat. Genet*. **9**: 407.

Tuiskula M. H., de Koning D.-J., Honkatukia M., Schuman N. F., Maki-Tanila A. and Vilkki J. 2004. Quantitative trait loci with parent-of-origin effects in chicken. *Genet. Res. Camb*. **84**: 57.

Vafiadis P., Bennett S. T., Todd J. A., Grabs R. and Polychronakos C. 1998. Divergence between genetic determinants of *IGF2* transcription levels in leukocytes and of *IDDM2*-encoded susceptibility to type 1 diabetes. *J. Clin. Endocrinol. Metab*. **83**: 2933.

Vafiadis P., Bennett S. T., Todd J. A., Nadeau J., Grabs R., Goodyer C. G., Wickramasinghe S., Colle E. and Polychronakos C. 1997. Insulin expression in human thymus is modulated by *INS* VNTR alleles at the *IDDM2* locus. *Nat. Genet.* **15**: 289.

Van Laere A. S., Nguyen M., Braunschweig M., Nezer C., Collette C., Moreau[3] L., Archibald A. L., Haley C. S., Buys N., Tally M., Andersson G., Georges M. and Andersson L. 2003. A regulatory mutation in *IGF2* causes a major QTL effect on muscle growth in the pig. *Nature* **425**: 832.

Visscher P. M., Thompson R. and Haley C. S. 1996. Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**: 1013.

Walling G. A. *et al.* 1998. Mapping of quantitative trait loci on porcine chromosome 4. *Anim. Genet.* **29**: 415.

Webber A. L., Ingram R. S., Levorse J. M. and Tilghman S. M. 1998. Location of enhancers is essential for the imprinting of *H19* and *IGF2* genes. *Nature* **391**: 711.

Weber M., Milligan L., Delalbre L., Antoine E., Brunel C., Cathala G. and Forne T. 2001. Extensive tissue-specific variation of allelic methylation in the *IGF2* gene during mouse fetal development: relation to expression and imprinting. *Mechanisms of Development* **101**: 133.

Weksberg R., Smith A. C., Squire J. and Sadowski P. 2003. Beckwith–Wiedemann syndrome demonstrates a role for epigenetic control of normal development. *Human Molecular Genetics* **12**: 61.

Wilson E. M. and Rotwein P. 2006. Control of MyoD function during initiation of muscle differentiation by an autocrine signaling pathway activated by insulin-like growth factor-II. *J Biol Chem*. **281**: 29962.

Winter A., Kramer W., Werner F. A., Kollers S., Kata S., et al. 2002. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (*DGAT1*) with variation at a quantitative trait locus for milk fat content. *Proc. Natl. Acad. Sci. USA* **99**: 9300.

Wylie A. A., Murphy S. K., Orton T. C. and Jirtle R. L. 2000. Novel imprinted *DLK1/GTL2* domain on human Chr 14 contains motifs that mimic those implicated in *IGF2/H19* regulation. *Genome Res*. **10**: 1711.

Yang G. C., Ren J., Guo Y. M., Ding N. S., Chen C. Y. and Huang L. S. 2006. Genetic evidence for the origin of an *IGF2* quantitative trait nucleotide in Chinese pigs. *Anim. Genet.* **37**: 179.

Zemel S., Bartolomei M. S. and Tilghman S. M. 1992. Physical linkage of two mammalian imprinted genes, *H19* and insulin-like growth factor 2. *Nat. Genet*. **2**: 61.

Resume

Minh Nguyen
Tel. 00.32.474.668.467
bambo_hn@yahoo.com
Rue de la Cote d'Or 167 – 4000 – Liège – BE

I. INTERESTS: Molecular Biology, Molecular Genetics, Epigenetics, Antibody, Vaccine

II. EDUCATIONS:

1998-2009: **Ph.D. in Veterinary Science (Genetics)** – University of Liège, Belgium
2006-2007: **Project Management** – Boston University Brussels: 8 credits
1996-1998: **M.Sc. in Molecular Biology -** KUL, VUB, UA, Belgium: 120 credits
1987-1992: **Doctor of Veterinary Medicine** - Hanoi University of Agriculture, Vietnam

III. EMPLOYMENTS:

2003-2006: **Scientist I**, Dyax sa, Liège, Belgium
1993-1996: **Researcher**, Dept. of Biochemistry and Immunology, National Institute of Veterinary Research, Hanoi, Vietnam

IV. TECHNICAL EXPERTISE AND EXPERIENCES:

1. Positional Cloning
2. Phage Display Technology
3. High Throughput (HT)-PCR, HT- Sanger Sequencing, Bisulphite Sequencing
4. DNA Cloning, RT-PCR, 5'-3'- RACE, Southern Blot, Northern Blot, Genotyping
5. Yeast Two Hybrid System, HIA, ELISA, RIA, fermentation system for vaccine production
6. Semi-automated system (WebPhage, Tecan and Minitrak) for HT-PCR, HT-Sequencing
7. Vector Nti 9.0, Sequencher, Bio-computing with Belgian EMBnet Node, & other databases
8. Phred-Phrap-Consed Package, Alfresco

V. AWARDS AND SCHOLARSHIPS:

1998-2003: Scholarship - University of Liege - Doctorate in Veterinary Science
1996-1998: Scholarship - ABOS - Master of Science in Molecular Biology
1987-1992: Golden medal in "National awards for students in scientific research"

VI. LANGUAGES:

Vietnamese: Native *English*: Excellent *French*: Elementary

VII. PUBLICATIONS:

1. Schoonbroodt S, Steukers M, Viswanathan M, Frans N, Timmermans M, Wehnert A, **Nguyen M**, Ladner RC, Hoet RM. 2008. Engineering antibody heavy chain CDR3 to create a phage display Fab library rich in antibodies that bind charged carbohydrates. ***J Immunol.*** **181**: 6213.
2. Anne.-Sophie Van Laere[1,*], **Minh Nguyen**[1,*], Martin Braunschweig[1], Carine Nezer[2], Catherine Collette[2], Laurence Moreau[2], Alan L. Archibald[3], Chris S. Haley[3], Nadine Buys[4], Göran Andersson[1], Michel Georges[2,#], and Leif Andersson[1,#]. 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. ***Nature*** **425**: 832.
3. Valerie Amarger[2,3*], **Minh Nguyen**[1*] , Anne-Sophie Van-Laere[2], Carine Nezer[1], Michel Georges[1] and Leif Andersson[2]. 2002. Comparative sequence analysis of Insulin, IGF2 and H19 gene cluster in pigs. ***Mamm. Genome*** **13**: 388.
4. **Minh Nguyen** & Phan Thanh Phuong, 1996. Study on immunosuppression induced by Trypanosoma evansi in buffaloes and its influence on the immune response to Haemorhagic Septicemia. International Workshop on diagnosis and control of Haemorhagic Septicemia. Bali, Indonesia.
5. **Minh Nguyen** et al. 1995. Preparation of the bivalent oil-adjuvant vaccine and its application. ***Research in Agriculture Biotechnology***: 358.
6. **Minh Nguyen** et al. 1995. Study on immune efficiency of adjuvant vaccine against Pasteurellosis for cattle and pigs. ***Research in Agriculture Biotechnology***: 362.