

**Implementation of SNPs in pig genetics:**  
**LD and QTL analysis**

Eerste promotor:

Prof. dr. M.A.M. Groenen

*Persoonlijk hoogleraar bij de leerstoelgroep Fokkerij en Genetica  
Wageningen Universiteit*

Tweede promotor:

Prof. dr. B.A. van Oost

*Hoogleraar Moleculaire Genetica bij de faculteit der Diergeneeskunde  
Universiteit Utrecht*

Co-promotor:

Dr. M.W.F. te Pas

*Senior Onderzoeker bij de Animal Sciences Group  
Wageningen Universiteit en Research Centrum*

Promotiecommissie:

Dr. W. Coppieters

*Université de Liège*

Prof. dr. J.A.M. Leunissen

*Wageningen Universiteit*

Prof. dr. P. Stam

*Wageningen Universiteit*

Prof. dr. P.J. de Wit

*Wageningen Universiteit*

Dit onderzoek is uitgevoerd binnen de onderzoeksschool WIAS.

# **Implementation of SNPs in pig genetics:**

## **LD and QTL analysis**

Bart Johan Jungerius

Proefschrift

Ter verkrijging van de graad van doctor  
op gezag van de rector magnificus  
van Wageningen Universiteit,  
prof.dr.ir. L. Speelman,  
In het openbaar te verdedigen  
op woensdag 27 oktober 2004  
des namiddags te vier uur in de Aula

B.J. Jungerius

Implementation of SNPs in pig genetics: LD and QTL analysis  
Thesis Wageningen University, The Netherlands, 2004  
-with summary in Dutch - 128 p

ISBN 90-8504-071-X

## **Abstract**

The aim of the work described in this thesis was the implementation and application of SNP markers in animal breeding and genetics. The emphasis was on the analysis of fatness traits in pigs, in particular of the imprinted QTL region on SSC2p. The identification of SNP markers in this region is described in Chapter 2. Chapters 3 and 4 describe POSA and SNPtyper, both are software tools that contribute to automation of data analysis in SNP discovery and SNP genotyping. The QTL on SSC2p was reanalysed after the identification of a SNP in the IGF2 gene that had high impact on lean muscle growth and back fat thickness. In Chapter 5 this re-analysis is described and it is concluded that the QTL is reduced to a quantitative trait nucleotide or QTN. For research in other genomic regions, the SNP marker density needed in association studies was an issue of discussion. To address this question, Chapter 6 describes an estimation of the extent of LD in the porcine genome. Although an average extent of LD was estimated, further investigation of LD and its extent is crucial in order to develop a map of the patterns of sequence variation in the porcine genome.



## **Table of Contents**

<b>Chapter 1</b>	General Introduction	1
<b>Chapter 2</b>	Development of a SNP map of porcine chromosome 2	15
<b>Chapter 3</b>	POSA, Perl objects for sequence analysis	33
<b>Chapter 4</b>	Typing single nucleotide polymorphisms using a gel-based sequencer: a new data analysis tool and suggestions for improved efficiency	43
<b>Chapter 5</b>	The IGF2-intron3-G3072A substitution completely explains a major imprinted QTL effect on backfat thickness in a Meishan X European White intercross	51
<b>Chapter 6</b>	Estimation of the extent of linkage disequilibrium in seven regions of the porcine genome	65
<b>Chapter 7</b>	General discussion	83
	Summary	101
	Samenvatting	107
	List of publications	113
	Nawoord	117
	Curriculum Vitae	121
	Training and supervision plan WIAS	125





# **Chapter 1**

## **General Introduction**



Over the last decades, there has been an enormous increase in knowledge of the genomes of livestock species. Advances in molecular biology techniques and the rapid development of DNA markers facilitated the development of comprehensive linkage maps of several livestock species, including pigs (Ellegren *et al.*, 1994; Archibald *et al.*, 1995; Rohrer *et al.*, 1994 and 1996).

In addition, the human genome sequencing project (The International Human Genome Sequencing Consortium, 2001) had high impact on animal genetic research and the whole genome sequences of several livestock species will become available in near future, starting with chicken (The Chicken Genome Consortium, 2004, in preparation).

Genetic maps of livestock genomes have been applied in several linkage studies to map loci and genes that underlie genetic variance of economically important traits. In pig production, fatness is an important trait that has been extensively studied over the last years.

### **Fatness traits in pigs**

In general, fat is deposited in four different depots in the body: subcutaneous (backfat), visceral, intermuscular and intramuscular fat (Kouba *et al.*, 1999). In pig production the two depots of highest interest are backfat (which accounts for 60-70% of total body fat) and intramuscular fat. In breeding programs, selection is directed against backfat thickness (BFT), because reduced carcass fatness benefits carcass quality and production efficiency. A reduction of overall carcass fatness, however, also reduces the level of intramuscular fat (IMF) and the correlated quality traits such as tenderness, juiciness and taste (Hovenier, 1993).

Identification of genes involved in intramuscular or backfat development could lead to a better understanding of the underlying genetic mechanisms and this could facilitate an implementation in breeding programs for reduced BFT without negative consequences for IMF.

### **Quantitative Trait Loci for fatness traits in pigs**

Fatness traits (like BFT and IMF) typically show a continuous distribution of phenotypic values rather than discrete phenotypic values and are therefore referred to as quantitative traits. Like most quantitative traits, BFT and IMF are controlled by multiple genes and may be influenced by several environmental factors. Genes influencing quantitative traits can be identified through two different strategies: the candidate gene approach and the whole genome scan approach.

The candidate gene approach applies previous knowledge about the functions of genes to select those genes that might be involved in the trait of interest. The selected candidate genes are tested for association with the trait or phenotype. The candidate gene approach has particularly been successful for relatively simple traits with only few

genes involved, like the identification of genes involved in coat color in pigs (Kijas *et al.*, 1998 and 2001).

In the whole genome scan approach, a large number of genetic markers from across the genome is analysed to observe segregation of chromosomal segments through a pedigree. By analyzing the (co-)segregation of the phenotypic trait or value with certain chromosomal segments in an experimental population, chromosomal regions harbouring genes affecting a quantitative trait can be identified (Geldermann *et al.*, 1996). Such a chromosomal region harbouring genes that influence the variation of a quantitative trait is referred to as a quantitative trait locus (QTL) (Andersson *et al.*, 1994).

In general, experimental populations for QTL mapping are comprised of intercrosses of phenotypically divergent founder populations. The assumption that the genes influencing the trait of interest are fixed in these founder populations enables identification of QTL regions explaining the genetic differences between populations.

In several studies, different experimental crosses have been used to detect QTL regions for fatness and meat quality traits in pigs (reviewed by Bidanel and Rothschild, 2002).

For BFT QTLs were identified on SSC1, 4 and 7 in several experimental populations (Andersson *et al.*, 1994; Knott *et al.*, 1998; Rohrer and Keele, 1998; Walling *et al.*, 1998; Wang *et al.*, 1998; de Koning *et al.*, 1999; Rohrer *et al.*, 2000; Pérez-Enciso *et al.*, 2000; Wada *et al.*, 2001; Bidanel *et al.*, 2001; Grindflek *et al.*, 2001; Malek *et al.*, 2001).

In addition, a paternally expressed (also referred to as maternally imprinted) QTL was identified on the p-arm of SSC2 in intercrosses between Large White and Wild-Boar (Jeon *et al.*, 1999), Large White and Pietrain (Nezer *et al.*, 1999) and European White and Meishan (de Koning *et al.*, 1999 and 2000; Rattink *et al.*, 2000).

For most detected QTL, the identified chromosomal regions are still large and harbour hundreds of genes. To narrow down these regions to the gene(s) or eventually the mutation causing the effect on the phenotype fine-mapping of this region is necessary. Fine-mapping can be achieved by adding more information, for example by typing additional markers in the region, by typing additional individuals or by applying additional approaches like linkage disequilibrium mapping and haplotype-sharing analysis.

### Imprinting

Genomic imprinting affects several dozen mammalian genes and results in the expression of those genes from only one of the two parental chromosomes. (Reik and Walter, 2001). Imprinting is an epigenetic activity: it is a gene-regulating activity that does not involve changes to the DNA sequence and that can persist through one or more generations (Pennisi, 2001). Although the mechanism behind imprinting is still not completely understood, methylation of the DNA seems to play a key role (Li *et al.*, 1993). Methylation patterns can be passed on to next generations and are known to

affect protein-DNA interaction and in general repress gene activity (Razin and Shemer, 1999).

Paternal expression (i.e. only the allele inherited from the paternal side is expressed and the allele from the maternal side is silenced) was first demonstrated for the murine insulin-like growth factor II (IGF2; DeChiara et al., 1991). In human and pig, IGF2 is also paternally expressed (Morison et al., 1998; Nezer et al., 1999).

In QTL analyses for BFT on SSC2, the analysis that only considered the paternal allele, gained higher significance levels than analyses focussed on maternal or both alleles. This suggests that the underlying gene also shows paternal expression rather than standard Mendelian expression. The QTL region on SSC2 is homologous to human chromosome 11p15 (HSA11p15) (Rattink et al., 2001). In human, this region contains several genes that are known to be imprinted. One of these genes is *IGF2*, which is involved in development and growth. In pigs, *IGF2* is located within the QTL region at the distal tip of the p-arm of SSC2 (Jeon et al., 1999). Regarding the location and the paternal expression *IGF2* gene is a good candidate for being the gene harbouring the mutation that underlies the QTL.

Recently, a single nucleotide polymorphism (SNP) in a regulatory element of the *IGF2* gene was shown to be correlated with the major QTL effect on muscle growth and BFT in crosses between Large White and Wild Boar or Pietrain. The SNP is a G-A substitution at position 3072 in the third intron of *IGF2*. Strong evidence for a causal relation with the observed QTL effect was provided (Van Laere et al., 2003). The influence of the *IGF2* mutation on the paternally expressed QTL for BFT in the cross between Meishan X European Whites will be explored in Chapter 5 of this thesis.

### **Single Nucleotide Polymorphisms**

According to Brookes (1999), the definition of a single nucleotide polymorphism is a single base pair position in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater. This definition is rather strict, as it holds some factors that exclude polymorphisms based on the allele frequencies in the population studied. In addition, insertion/deletion polymorphisms (indels) are excluded, since they are created through a different mechanism and do not have an alternative nucleotide on the polymorphic position. In practice, all subtle sequence differences, including indels, are potential markers for mapping purposes, even those with low frequencies. In addition, the frequency is dependent of the studied population. Consequently, a polymorphism might be a true SNP in one population and just a subtle DNA difference in another. Altogether, it is neither practicle nor useful to classify all subtle DNA differences as either true SNPs or indels or 'low'frequency polymorphisms. Most polymorphisms can be used for the same goal(s), using the same methods, regardless

of allele frequencies, population(s) or origin of the mutation. For these reasons, in this thesis a less strict definition of SNPs is applied and all subtle DNA sequence differences are considered SNPs, as is commonly done in practice.

In practice, SNPs are predominantly bi-allelic, even though in principle any of the four nucleotides can be present at any position in a stretch of DNA. This is due to the low frequency of mutations that leads to new SNPs. This mutation rate is estimated at  $10^{-8}$  changes per nucleotide per generation (Crow, 1995; Li et al., 1996) and corresponds to about 100 new SNPs per individual (Kondrashov, 1995; Crow, 1995). As a result, the chance for a second independent mutation that introduces a third allele at the same position is very low.

Basically, the bi-allelic SNPs are comprised of two different categories: transitions and transversions. In transitions, a purine is exchanged for the other purine ( $A \rightleftharpoons G$ ), while on the reverse strand a pyrimidine is exchanged for the other pyrimidine ( $C \rightleftharpoons T$ ). Transversions consist of purine-pyrimidine (and their complementary pyrimidine-purine exchanges):  $A \rightleftharpoons C$  ( $T \rightleftharpoons G$ ),  $A \rightleftharpoons T$  ( $T \rightleftharpoons A$ ),  $C \rightleftharpoons G$  ( $G \rightleftharpoons C$ ). The occurrences of transitions and transversions are not equal in the genome. Although there are three times as many options for transversions as for transitions, transitions are found twice as often as transversions. The higher level of transitions might be related to 5-methylcytosine (5mC) deamination reactions. 5mC is the product of post-synthetic modification of cytosine residues and occurs primarily in CpG doublets in the mammalian genome. 5mC is a mutable site that can undergo spontaneous deamination to thymine. Although a repair mechanism specifically recognizes G-T mismatches and replaces the thymine with cytosine, the 5mC to T transition mutation occurs about 10 times more often than other transitions (Holliday and Grigg, 1993).

The rate at which nucleotide differences are observed between two randomly chosen chromosomes is called the nucleotide diversity index (Nei and Li, 1979). In the human genome, this index is estimated at one in a thousand basepairs (Li and Sadler, 1991). Screening more chromosomes (more individuals) will identify more polymorphisms, but the nucleotide diversity index remains constant and thus allows comparisons between studies applying different sample sizes. Compared to the total human genome, the nucleotide diversity index in coding regions is about 4-fold lower and about half of these coding mutations result in non-synonymous codon changes (Li and Sadler, 1991; Nickerson 1998).

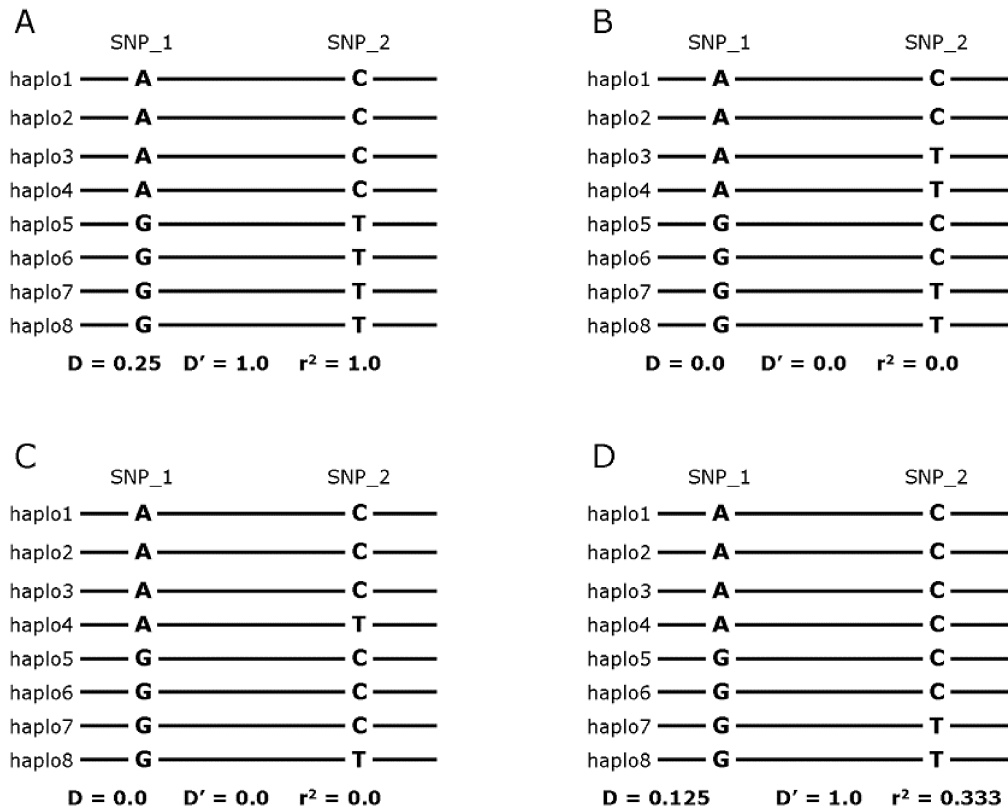
Nucleotide diversity indexes are reported to be 1/1331 bp in humans (Sachidanandam et al., 2001), 1/443 bp in cattle (Heaton et al., 2001), 1/515 bp in mice (Lindblad-Toh et al., 2000). Recently, Fahrenkrug et al. (2002) reported porcine SNP densities that translate in an index of 1/609 bp. Overall, this adds up to about several millions of SNPs between any two individuals and about 100000 amino acid changes in the proteomes. It has been estimated that in the world's human population, about 10 million SNPs (that is, 1 variant

site per 300 bases on average) vary such that both alleles are observed at a frequency of 1% or greater, and that these 10 million common SNPs constitute 90 % of the variation in the population (Kruglyak and Nickerson, 2001; Reich et al., 2003). The remaining 10% is due to a vast array of variants that are each rare in the population. Nearly all variant sites result from a single historical event as the mutation rate is low ( $10^{-8}$  per site per generation) compared to the number of generations since the most recent common ancestor of any two humans (of the order of  $10^4$  generations).

### **Linkage Disequilibrium**

Particular alleles at neighbouring loci tend to be co-inherited. For tightly linked loci, this might lead to associations between alleles in a population. The result is a non-random pattern of association between alleles at different genetic loci. This (statistical) association of sequence variants at different loci along the chromosome is called linkage disequilibrium (LD).

Several measures to quantify LD exist and most of them are based on Lewontin's  $D$  (reviewed by Devlin and Risch, 1995). Lewontin's  $D$  considers two loci and quantifies the difference between the observed frequency of two particular alleles occurring together and the expected frequency based on the single allele frequencies.  $D'$  is frequently used and it is formed by dividing  $D$  by the maximum value of  $D$  with regard to the allele frequencies.  $D'$  has the same range of values regardless of the frequencies of the SNPs compared and varies between 0 (complete linkage equilibrium; LE) and 1 (complete LD). Intermediate values of  $D'$ , however, have no clear interpretation. Recently, an alternative measure of LD has emerged as the measure of choice for quantifying and comparing LD in the context of mapping:  $r^2$ , sometimes denoted as  $\rho^2$  (Pritchard et al., 2001; Weiss and Clark, 2002). The value of  $r^2$  reflects the correlation of alleles at two loci and is formed by dividing the square of  $D$  by the product of the four allele frequencies. As with  $D'$ , a  $r^2$  value of 1 corresponds to complete LD and a value of 0 corresponds with complete LE. In contrast to  $D'$ , intermediate values of  $r^2$  can be interpreted: Consider a two locus model where locus 1 is functionally associated with phenotype and nearby locus 2 is in LD with locus 1 with a  $r^2$  value of .33. In order to have the same power to detect association between the phenotype and locus 2, the sample size needs to be expanded with a factor 3 ( $=1/r^2$ ). The choice of the LD measurement used has an impact on the results and, consequently, on the conclusion, as is demonstrated in Figure 1.



**Figure 1:** Haplotypes consisting of two SNP loci from different individuals.

A: the A allele of SNP1 always coincides with the G on SNP2. Both SNPs are in complete LD. Consequently, both  $|D'|$  and  $r^2$  are equal to 1. B: the allelic combinations appear to be random and the SNPs are in linkage equilibrium (LE), as reflected by both measures of LD, which both equal 0. C: a random distribution of alleles with non-equal allele frequencies. Again, the markers are in LE. D: the allele frequencies are non-equally distributed and not random: only three of the four possible combinations occur.  $D'$  reflect complete LD, whereas  $r^2$  reflects only partial LD.

Typically, LD is the result of a mutation that introduces a newly created allele surrounded by a series of alleles at other polymorphic loci. Such a specific group of alleles present on a single chromosome is called a haplotype. Thus, at the time of mutation, a new unique group of alleles (haplotype) is established and it can be passed on to future generations. In this new haplotype the new allele is completely predictive of its flanking alleles and therefore it is in complete LD with the surrounding alleles.

Reversibly, newly established LD can be disturbed by meiotic recombination events. In these events portions of sister chromosomes are exchanged and this leads to 'random' shuffling of alleles at different loci along the chromosome. As a result, LD will erode with



the number of generations since the haplotype was established and will thus erode with time. the chance of alleles to be in LD is higher for loci closer together, because the shorter the distance between two loci the lower the chance of recombination to disturb the haplotype.

Recombination rates do not seem to be equally distributed over the chromosome as some recombination events occur repeatedly at 'hotspots' (Jeffreys et al., 2001; Chakravarti et al., 1984). As a result of hotspot-oriented recombination, today's chromosomes comprise a mosaic of haplotype blocks derived from ancestral chromosome fragments (Cargill et al., 1999) and shared discrete haplotype blocks and LD patterns can be observed even in apparently unrelated individuals and populations.

In human data, simulations and empirical data have suggested that LD may extend for somewhere between 5 and 500 kb, that the extent of LD is not equally distributed over the genome (Clark et al., 1998; Rieder et al., 1999; Moffatt et al., 2000; Templeton et al., 2000) and that LD patterns vary between populations (Goddard et al., 2000; Kidd et al., 2000; Reich et al., 2000).

Despite the recent developments in LD research in human, there is only little information available on LD or the the extent of LD in pigs. A better knowledge of LD in the porcine genome is of high relevance for pig genome research as well as pig breeding.

### **Aim and outline of this thesis**

The aim of the work described in this thesis is the implementation and application of SNP markers in animal breeding and genetics. The emphasis is on analysis of fatness traits in pigs with special attention for the imprinted QTL region on SSC2p. A number of SNPs is identified in this QTL region by checking alignments of sequences of PCR products from a panel of individuals with different genetic backgrounds, as described in **Chapter 2**. In **Chapter 3**, a tool, POSA, to enable easy and flexible sequence analysis is described. POSA is developed to reduce the time spent on manual editing of sequence contigs. In **Chapter 4** another tool, SNPtyper, is described. SNPtyper offers an alternative to analyse multiplex SBE reactions on a gel-based sequencer like the ABI377. In **Chapters 5** and **6** the emphasis is on the applications and consequences of SNPs in the porcine genome. **Chapter 5** describes reanalyses of the paternally expressed QTL on SSC2 with the inclusion of a functional SNP in the IGF2 gene. In **Chapter 6**, SNPs are applied in an estimation of the extent of LD in pigs. Finally, in **Chapter 7** the results of this thesis are discussed and the implications of this work are explored.

## References

- Andersson L, Haley CS, Ellegren H, Knott SA, Johansson M, Andersson K, Andersson-Eklund L, Edfors-Lilja I, Fredholm M, Hansson I, et al. (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science*. 263(5154):1771-4.
- Archibald AL, Haley CS, Brown JF, Couperwhite S, McQueen HA, Nicholson D, Coppieters W, Van de Weghe A, Stratil A, Wintero AK, et al. (1995) The PiGMaP consortium linkage map of the pig (*Sus scrofa*). *Mamm Genome* 6(3):157-75.
- Bidanel JP, Milan D, Iannuccelli N, Amigues Y, Boscher MY, Bourgeois F, Caritez JC, Gruand J, Le Roy P, Lagant H, Quintanilla R, Renard C, Gellin J, Ollivier L, Chevalet C. (2001) Detection of quantitative trait loci for growth and fatness in pigs. *Genet Sel Evol* 33(3):289-309.
- Bidanel JP, Rothschild M. (2002) Current status of quantitative trait locus mapping in pigs. *Pig News and Information* 23(2): 39N-45N
- Brookes AJ. (1999) The essence of SNPs. *Gene* 234(2):177-86.
- Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*. 22(3):231-8.
- Chakravarti A. (1999) Population genetics--making sense out of sequence. *Nat Genet*. 21(1 Suppl):56-60.
- Chakravarti A, Buetow KH, Antonarakis SE, Waber PG, Boehm CD, Kazazian HH. (1984) Nonuniform recombination within the human beta-globin gene cluster. *Am J Hum Genet*. 36(6):1239-58.
- Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet*. 63(2):595-612.
- Crow JF. (1995) Spontaneous mutation as a risk factor. *Exp Clin Immunogenet*. 12(3):121-8.
- de Koning DJ, Rattink AP, Harlizius B, van Arendonk JA, Brascamp EW, Groenen MA. (2000) Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proc Natl Acad Sci U S A*. 97(14):7947-50.

de Koning DJ, Janss LL, Rattink AP, van Oers PA, de Vries BJ, Groenen MA, van der Poel JJ, de Groot PN, Brascamp EW, van Arendonk JA. (1999) Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (*Sus scrofa*). *Genetics*. 152(4):1679-90.

DeChiara TM, Robertson EJ, Efstratiadis A. (1991) Parental imprinting of the mouse insulin-like growth factor II gene. *Cell* 64(4):849-59.

Devlin B, Risch N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29(2):311-22.

Ellegren H, Chowdhary BP, Johansson M, Marklund L, Fredholm M, Gustavsson I, Andersson L. (1994) A primary linkage map of the porcine genome reveals a low rate of genetic recombination. *Genetics* 137(4):1089-100.

Fahrenkrug SC, Freking BA, Smith TP, Rohrer GA, Keele JW. (2002) Single nucleotide polymorphism (SNP) discovery in porcine expressed genes. *Anim Genet*. 33(3):186-95.

Geldermann H, Muller E, Beeckmann P, Knorr C, Yue G, Moser G. (1996) Mapping of quantitative-trait loci by means of marker genes in F2 generations of Wild boar, Pietrain and Meishan pigs. *J Anim Breed Genet*. 113: 381-7.

Goddard KA, Hopkins PJ, Hall JM, Witte JS. (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet*. 66(1):216-34.

Grindflek E, Szyda J, Liu Z, Lien S. (2001) Detection of quantitative trait loci for meat quality in a commercial slaughter pig cross. *Mamm Genome* 12(4):299-304.

Heaton MP, Grosse WM, Kappes SM, Keele JW, Chitko-McKown CG, Cundiff LV, Braun A, Little DP, Laegreid WW. (2001) Estimation of DNA sequence diversity in bovine cytokine genes. *Mamm Genome* 12(1):32-7.

Hovenier R. (1993) Breeding for meat quality in pigs. Thesis Wageningen Agricultural University.

Jeffreys AJ, Kauppi L, Neumann R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*. 29(2):217-22.

Jeon JT, Carlborg O, Tornsten A, Giuffra E, Amarger V, Chardon P, Andersson-Eklund L, Andersson K, Hansson I, Lundstrom K, Andersson L. (1999) A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nat Genet*. 21(2):157-8.

Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK. (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am J Hum Genet.* 66(6):1882-99.

Kijas JM, Moller M, Plastow G, Andersson L. (2001) A frameshift mutation in MC1R and a high frequency of somatic reversions cause black spotting in pigs. *Genetics* 158(2):779-85.

Kijas JM, Wales R, Tornsten A, Chardon P, Moller M, Andersson L. (1998) Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. *Genetics* 150(3):1177-85.

Knott SA, Marklund L, Haley CS, Andersson K, Davies W, Ellegren H, Fredholm M, Hansson I, Hoyheim B, Lundstrom K, Moller M, Andersson L. (1998) Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* 149(2):1069-80.

Kondrashov AS. (1995) Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J Theor Biol.* 175(4):583-94.

Kouba M, Bonneau M, Noblet J. (1999) Relative development of subcutaneous, intermuscular, and kidney fat in growing pigs with different body compositions. *J Anim Sci.* 77(3):622-9.

Kruglyak L, Nickerson DA. (2001) Variation is the spice of life. *Nat Genet.* 27(3):234-6.

Li E, Beard C, Jaenisch R. (1993) Role for DNA methylation in genomic imprinting. *Nature* 366(6453):362-5.

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol.* 5(1):182-7.

Li WH, Sadler LA. (1991) Low nucleotide diversity in man. *Genetics* 129(2):513-23.

Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette JP, Ardlie K, Reich DE, Robinson E, Sklar P, Shah N, Thomas D, Fan JB, Gingeras T, Warrington J, Patil N, Hudson TJ, Lander ES. (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet.* 24(4):381-6.

Malek M, Dekkers JC, Lee HK, Baas TJ, Rothschild MF. (2001) A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. I. Growth and body composition. *Mamm Genome* 12(8):630-6.

Moffatt MF, Traherne JA, Abecasis GR, Cookson WO. (2000) Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. *Hum Mol Genet.* 9(7):1011-9.

Morison IM, Reeve AE. (1998) A catalogue of imprinted genes and parent-of-origin effects in humans and animals. *Hum Mol Genet.* 7(10):1599-609.

Nei M, Li WH. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A.* 76(10):5269-73.

Nezer C, Moreau L, Brouwers B, Coppieters W, Dettleux J, Hanset R, Karim L, Kvasz A, Leroy P, 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(2):155-6.

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet.* 19(3):233-40.

Pennisi E. (2001) Behind the scenes of gene expression. *Science* 293(5532):1064-7.

Perez-Enciso M, Clop A, Noguera JL, Ovilo C, Coll A, Folch JM, Babot D, Estany J, Oliver MA, Diaz I, Sanchez A. (2000) A QTL on pig chromosome 4 affects fatty acid metabolism: evidence from an Iberian by Landrace intercross. *J Anim Sci.* 78(10):2525-31.

Rattink AP, Faivre M, Jungerius BJ, Groenen MA, Harlizius B. (2001) A high-resolution comparative RH map of porcine chromosome (SSC) 2. *Mamm Genome* 12(5):366-70.

Rattink AP, De Koning DJ, Faivre M, Harlizius B, van Arendonk JA, Groenen MA. (2000) Fine mapping and imprinting analysis for fatness trait QTLs in pigs. *Mamm Genome* 11(8):656-61.

Razin A, Shemer R. (1999) Epigenetic control of gene expression. *Results Probl Cell Differ.* 25:189-204.

Reich DE, Gabriel SB, Altshuler D. (2003) Quality and completeness of SNP databases. *Nat Genet.* 33(4):457-8.

Reich DE, Goldstein DB. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol.* 20(1):4-16.

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

Rieder MJ, Taylor SL, Clark AG, Nickerson DA. (1999) Sequence variation in the human angiotensin converting enzyme. *Nat Genet.* 22(1):59-62.

Rohrer GA. (2000) Identification of quantitative trait loci affecting birth characters and accumulation of backfat and weight in a Meishan-White Composite resource population. *J Anim Sci.* 78(10):2547-53.

Rohrer GA, Keele JW. (1998) Identification of quantitative trait loci affecting carcass composition in swine: I. Fat deposition traits. *J Anim Sci.* 76(9):2247-54.

Rohrer GA, Alexander LJ, Hu Z, Smith TP, Keele JW, Beattie CW. (1996) A comprehensive map of the porcine genome. *Genome Res.* 6(5):371-91.

Rohrer GA, Alexander LJ, Keele JW, Smith TP, Beattie CW. (1994) A microsatellite linkage map of the porcine genome. *Genetics* 136(1):231-45.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D; International SNP Map Working Group. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409(6822):928-33.

The International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF. (2000) Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am J Hum Genet.* 66(1):69-83.

Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, Andersson G, Georges M, Andersson L. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* 425(6960):832-6.

Wada Y, Akita T, Awata T, Furukawa T, Sugai N, Inage Y, Ishii K, Ito Y, Kobayashi E, Kusumoto H, Matsumoto T, Mikawa S, Miyake M, Murase A, Shimanuki S, Sugiyama T, Uchida Y, Yanai S, Yasue H. (2000) Quantitative trait loci (QTL) analysis in a Meishan x Gottingen cross population. *Anim Genet.* 31(6):376-84.

Walling GA, Archibald AL, Cattermole JA, Downing AC, Finlayson HA, Nicholson D, Visscher PM, Walker CA, Haley CS. (1998) Mapping of quantitative trait loci on porcine chromosome 4. *Anim Genet.* 29(6):415-24.

Wang L, Yu TP, Tuggle CK, Liu HC, Rothschild MF. (1998) A directed search for quantitative trait loci on chromosomes 4 and 7 in pigs. *J Anim Sci.* 76(10):2560-7.

## **Chapter 2**

### **Development of a single nucleotide polymorphism map of porcine chromosome 2**

Bart Jungerius<sup>1</sup>, Annemieke Rattink<sup>1,2</sup>, Richard Crooijmans<sup>1</sup>, Jan van der Poel<sup>1</sup>, Bernard van Oost<sup>3</sup>, Marinus te Pas<sup>4</sup> and Martien Groenen<sup>1</sup>

<sup>1</sup> Animal Breeding and Genetics Group, Wageningen University,  
Marijkeweg 40, 6709 PG Wageningen, The Netherlands

<sup>2</sup> Present address: Nutreco, Veerweg 38, 5830 AE Boxmeer, The Netherlands

<sup>3</sup> Department of Clinical Sciences of Companion Animals, Utrecht University,  
PO box 80154, 5308 TD Utrecht, The Netherlands

<sup>4</sup> ID-Lelystad, Institute for Animal Science and Health,  
PO box 65, 8200 AB Lelystad, the Netherlands

Animal Genetics 34: 429-437 (2003)

### Summary

Single nucleotide polymorphism (SNP) markers are developed on porcine chromosome 2 (SSC2), predominantly on the p-arm. Several studies reported a QTL for backfat thickness in this region. SNPs were identified by comparative re-sequencing of PCR products from a panel of eight individuals. The panel consisted of five Large Whites, each from a different Dutch breeding company, a Meishan, a Pietrain and a Wild Boar. In total, 67 different PCR products were sequenced and 301 SNPs were identified in 32429 bp of consensus sequence, an average of one SNP in every 108 bp. After correction for sample size, this polymorphism rate corresponds to a heterozygosity value of 1 SNP every 357 bp.

For 63% of the SNPs there was variation among the five Large Whites and these SNPs are relevant for linkage and association studies in commercial populations. Comparing the Whites with other breeds revealed higher variation rates with Meishan: 89%, Pietrain: 69%, Wild Boar: 70%. Because many of the experimental populations to identify QTL are based on crosses between these breeds, these SNPs are relevant for the fine mapping of the QTL identified within these crosses.



## Introduction

Total genome scans have been successfully applied to identify quantitative trait loci (QTL) affecting economically relevant traits in livestock. In pigs, several carcass and fatness traits have been studied. On the p-arm of porcine chromosome 2 (SSC2) a maternally imprinted QTL for backfat thickness (BFT) has been identified close to the *IGF2* gene (Nezer et al., 1999; De Koning et al., 1999; Rattink et al., 2000). The results of Rattink et al. (2000) indicate that more than one QTL for BFT might be located in this region. The identified region spans over 60 cM, which corresponds to chromosomal fragments containing hundreds of genes. Fine mapping of this large region is necessary to determine haplotypes associated with these effects and to identify the underlying genes responsible for the observed QTL effects. Although the porcine map has expanded gradually over the past years (Archibald et al., 1995; Rohrer et al., 1996; Hawken et al., 1999), more markers need to be developed to reach a sufficient density that is required to facilitate detailed haplotype analysis.

To date, the markers used in genome scans are almost exclusively microsatellite markers, which occur once every 30 to 46 kb in pigs (Winterø et al., 1992). SNP markers are far more abundant with an occurrence of about 1 SNP per kb in humans (Wang et al., 1998) and about 1 SNP per 500 bp in mice (Lindblad-Toh et al., 2000) and cattle (Heaton et al., 2001). Because of their abundance SNPs have high potential for detailed haplotype analysis and application in association studies. In addition, SNPs are more suitable for association studies because they are almost exclusively identical by descent because of their relatively low mutation rate.

A straightforward strategy for the identification of SNPs is locus specific amplification (LSA) and comparative re-sequencing from multiple individuals (Rieder et al., 1998). In this strategy, PCR products amplified from genomic DNA from different individuals are directly sequenced and compared. This strategy has the advantage that the required coverage of several reads per locus can be reached immediately and the generated sequence traces can be compared directly. In addition, sequencing of a PCR product amplified using primers designed from mapped sequences will yield SNPs with known map positions. The speed-limiting factor in the LSA strategy is the design of primers for amplification. Many SNP genotyping assays, however, require a step of amplification of the region flanking the SNP in which the amplification primers can be employed again. The LSA strategy facilitates targeting of specific regions of interest. Here, we applied this approach to increase marker density on SSC2.

## Materials and methods

### *Animals*

A SNP discovery panel consisting of eight animals each from a different breed was used. The panel consisted of one Meishan boar, one Pietrain boar, one Wild Boar and five Large White sows. The Large Whites each originate from a different Dutch breeding company: Bovar (presently Hypor), Fomeva, Dutch Pig Herd Book (presently part of Topigs), Dalland and Euribrid.

### *Primer design*

A total of 99 STSs were tested for amplification and sequencing. Primers were based on one of four different sources: a) Thirty-one primer pairs were previously described in the literature; b) Twenty-five primer pairs were designed from porcine genomic sequences present in the Genbank database; c) Five primer pairs were designed from end-sequences of BAC clones that contain one or more loci previously mapped to SSC2 (Rattink et al., 2001a); d) Thirty-eight primer pairs were designed from end-sequences of subclones derived from BACs that contain one or more loci previously mapped to SSC2 (Rattink et al., 2001a). All BAC clones are from the porcine BAC library by Rogel-Gaillard et al. (1999). All primers were designed using Primer3 through the web interface ([http://www-genome.wi.mit.edu/cgi-bin/primer/primer3\\_www.cgi](http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi)).

### *PCR and sequencing*

PCR reactions were carried out in 24 µl volumes. Reactions contained 180 ng DNA, 1.5 mM MgCl<sub>2</sub>, 50 mM KCl, 10 mM Tris-HCl pH 8.3, 1 mM tetramethylammoniumchloride (TMAC), 0.1% Triton X-100, 0.01% gelatine, 200 µM of each dNTP, 0.5 unit Silverstar DNA polymerase (Eurogentec, Seraign, Belgium), 6 pmoles of each primer, and were covered with 12 µl mineral oil. Reactions were performed in a PTC100 thermocycler (MJ Research, Watertown, MA, USA) with a 5 min initial denaturation at 95°C followed by 35 cycles of 30 s at 95°C, 45 s at annealing temperature and 60 s at 72°C, and a final extension for 10 min at 72°C. The annealing temperatures were optimised for each primer pair and varied between 50°C and 65°C (see table 1). After PCR the excess of primers was removed by running the samples over a column of BioRad P-100 (BioRad, Hercules, CA, USA) in a Multiscreen MAHV N45 plate (Millipore, Billerica, MA, USA) in 96-well plate format. From each sample, 2 µl was checked on agarose gel to estimate the DNA concentration.

Sequencing reactions were performed with either the forward or the reverse amplification primer. Cycle sequencing reactions contained 100-400 ng of purified PCR

product, 2 µl of Big Dye Terminator Rtmix (Perkin-Elmer, Foster City, CA, USA), 2 µl of Half Big Dye Buffer (Genetix, New Milton, UK) and 0.8 pmol of either primer in a final volume of 10 µl. Excess dye terminator was removed by running the samples over a column of Sephadex G-50 (Amersham Pharmacia, Uppsala, Sweden) in a Multiscreen MAHV N45 plate format. Subsequently, the samples were dried using a SpeedVac and analysed in a 48 well format for 7hrs on a 36 cm 4.75% denaturing Long Ranger gel (FMC, Rockland, ME, USA) on an automated sequencer (ABI377, ABI, Foster City, CA, USA).

### *Polymorphism identification*

Sequencing gel images were analysed using the Sequencing Analysis software (ABI) for lane tracking and trace file extraction. Subsequently, all trace files were analysed using the Pregap4 program of the Staden software package (Bonfield and Staden, 1996; <http://www.mrc-lmb.cam.ac.uk/pubseq>). Using Pregap4, all chromatograms were analysed: bases were called with Phred (Ewing and Green, 1998, Ewing *et al.*, 1998) and low quality regions were masked. Next, all sequences passing Pregap4 were entered in a Gap4 database. In the Gap4 program (Bonfield *et al.*, 1995) a normal shotgun assembly was performed with a minimal initial match of 20 bp, a maximum of 25 pads per read and a maximum of 5 % mismatches. If necessary, the assembly was finished manually using the Gap4 *Join contig* interface. Positions at which a disagreement occurred were highlighted in the Gap4 contig editor window using the *Highlight disagreements* option. All nucleotide positions at which a disagreement occurred were tagged as a putative SNP position.

Consensus sequences were submitted to Genbank and the corresponding accession numbers are presented in Table 1. In the consensus sequences, the polymorphisms identified are represented by the IUB ambiguity codes reflecting the alleles found in the discovery panel. Insertion/deletion polymorphisms (indels) are listed as features of the sequence and are not represented in the consensus sequence. At indel positions the consensus reflects the sequences with all insertions, thus representing the longest possible haplotype for the sequence.

**Table 1:** Porcine sequence tagged sites (STS) primers used for amplification and sequencing.

STS name	Accession number	Forward primer	Reverse primer	PCR temp. <sup>1</sup>	Product length (bp) <sup>2</sup>	Source Type <sup>3</sup>	Source Reference <sup>4</sup>	Locus <sup>5</sup>	# of SNPs	Contig length (bp) <sup>6</sup>
RNHsts1	BV012582	AGGCTCCTTTTCCTGGAC	GTGCCGAAAGTTCTCTCAGC	60	100	L	Rattink et al, 2001	RNH	1	100
MUC5ACsts1	BV012580	CGGTCCAGGTGACTGAATCT	GGAGCAGACCCCTTCTGTGAC	60	125	L	Rattink et al, 2001	MUC5AC	0	125
P009C08sts1	BV079402	TGGCAACTTTACCCTGGAAC	AACACCTAGTAGACCTGCTC	55	<b>1300</b>	S	0196B06	MUC5AC	6	1121
H19sts1	BV079393	AAGCTCCAGCGACACTGTTT	GAGCAGAATCTTCCAGAACT	55	326	L	Rattink et al, 2001	H19	2	312
IGF2sts1	BV079394	AATTCTGGGTGCCACCATC	ATACAGACCAAGCCAATTGG	60	213	L	Rattink et al, 2001	IGF2	0	210
DHCR7sts1	BV079386	GAATCGGGAAGTGGTTTGAC	TGCAGGATGTTGACCAAGAC	65	148	G	AW786254	DHCR7	1	148
GALsts1	BV079391	TTTATTTTGGCGCTCAAACC	ATCTTCGTCTCGGCGTTGT	60	150	L	Rattink et al, 2001	GAL	0	150
P006B10sts1	BV012574	GCCAGAATCCTTAGACTTGG	GATCTAGACTGACCACTGAC	55	678	S	0364A05	GAL	0	679
P006F10sts1	BV079370	AGGTTGTCGATACGTCAGTG	CCTAACTCTGATGTAATCCG	58	1249	S	0364A05	GAL	23	1249
BEQ0718H07	BV012570	GCTCAGAACTGACCCTTTGC	GGCAAGTTACCTCCCTCTC	58	221	B	0718H07	SW2623	0	221
BER0718H07	BV012572	TCTATCTGTGAGTCCCTCTGGTC	TGGTGTTAGTGGGTAAACCAAG	58	332	B	0718H07	SW2623	1	333
P005E09sts1	BV012573	TAAGTCACCTCCAGGCTGAG	GCTGTCTGTGACATTAGGC	58	958	S	0718H07	SW2623	20	958
P011C10sts1	BV079419	ATGCACCGCCTCTATTACCC	GTTCCAAAGGTTGAGACTCC	58	<b>1200</b>	S	0668D06	SW2623	11	1074
P008B07sts1	BV079403	AAGCAAGTCAGATGCGACACC	CTGCAACTATGAGTCGATCC	60	<b>1500</b>	S	0234B08	SW256	23	1109
P004A09sts1	BV079368	TCCGCACATTGATGTGACAG	GAAGTCTGAGCTGCATTAGG	58	713	S	0226A03	SWR783	1	713
ADRBK1sts1	BV012576	GCAGCAGGAGGTAGCAGAG	TGCCCATCTTGGACATGTAG	62	310	L	Lyons et al, 1997	ADRBK1	6	310
ADRBK1sts2	BV079376	GATCAGCTACGTGTACAAGAC	CCTCTCTGCAAGAACTCACG	55	369	L	Rattink et al, 2001	ADRBK1	2	349
ACTN3sts2	BV079373	CAAGAACTACATCACTGCTGAGG	CGTAGAGGGCACTGGAGAAG	65	146	G	AJ301019	ACTN3	5	146
CFL1sts1	BV079381	TTTAACGACACCCAGTTCC	CTGGTCCTGCTTCCATGAGT	55	216	L	Rattink et al, 2001	CFL1	6	216
P011C03sts1	BV079404	TCTGCATTTCCTTACCGGAC	CACGTCATCTCATGCAATCC	55	<b>4000</b>	S	0322F10	S0141	20	1203
CNTFsts1	BV079382	ACAGTTTCTCTGAGGCCTCAC	TCAGACGAGTCATCCAGAAC	60	263	L	Rattink et al, 2001	CNTF	1	261
CNTFsts2	BV079383	TGGCTAGCAAGGAAGATTCTG	GAGTGTGATTTCTGACTAGTAGGC	58	789	G	u57644	CNTF	5	790
CNTFsts3	BV079384	AAGGCCTGTGATGGAACAC	TCAGCTCACTCCATCGATCA	62	717	G	u57644	CNTF	7	718
TCN1sts1	BV012583	AAACTCAAACCCCTCCTCCGT	TGTGCAAAGAGAGAAGCCAA	62	134	L	Rattink et al, 2001	TCN1	1	134
FTH1sts1	BV079390	TAAGCTGGCCTCCCGGAGAC	GGTACACTAAGGAAAGAACT	55	130	L	Rattink et al, 2001	FTH1	2	130
LHX1sts1	BV079395	TGGGCCTCTAAAGGACACAG	CTGGGAAATGGAAGCACAGT	62	349	G	af063245	LHX1 / FTH1	1	332
P006A09sts1	BV079369	AACCAACATTACTGCCTGGCA	AGCTCACAAACAGGCTTCTC	60	662	S	0978F05	FTH1	18	663

STS name	Accession number	Forward primer	Reverse primer	PCR temp. <sup>1</sup>	Product length (bp) <sup>2</sup>	Source Type <sup>3</sup>	Source Reference <sup>4</sup>	Locus <sup>5</sup>	# of SNPs	Contig length (bp) <sup>6</sup>
P006H08sts1	BV012575	TCCAAGTAGCACATCAAAGC	CTGACATTGAGATTTCATCC	55	347	S	0978F05	FTH1	3	348
PLCB3sts2	BV079406	TGAGTCGGTCAACTCCATCC	GCTGCTCCTGACACTCCTG	62	900	G	BG835750	PLCB3	6	498
MEN1sts1	BV079407	GACCTCTCATCCGACCCCTTT	TCATCTTCCTCGCAACTGAA	62	152	G	BG732325	MEN1	0	110
P011C12sts1	BV079416	CACGTCATCTCATGCAATCC	TCTGCATTTCCTTACCGGAC	55	4000	S	0783A10	SWR1445	21	1082
P011B02sts1	BV079413	TATCTGGATGGATGTCTCAC	AGAAGTGTATCATGCGGACC	55	464	S	0254F09	SW1450	0	464
P011G01sts1	BV079414	CAACAAGCCACGTGTTCTGTG	TATCCTTCTGTAGCTACCTG	55	4000	S	0254F09	SW1450	7	747
CD59sts2	BV079380	TGCCACCCAAAACCTTACTACC	GGCCTGGTCTTCAAAGTCGCT	55	378	G	af020302	CD59	2	377
CATsts1	BV079378	TGCCTCTGAAACAAAACGTG	TTCAAAAGACCCCAAAGCAT	55	458	L	Rattink et al, 2001	CAT	1	448
CATsts2	BV079379	GCTGAGTAACCCAGAGAATGC	GCAATTACAAATCAGTCTGTTGC	62	702	G	d89812	CAT	1	702
WT1sts1	BV079371	TTAACATTCTCTGGCTCG	GCCTTGCCCTCTGATTTATTT	60	425	L	Rattink et al, 2001	WT1	2	425
FSHBsts1	BV079408	CATTGCCATGAGCTATGGTG	TCCTTGACCTATCACGAGGC	62	381	L	Moran, 1993	FSHB	0	195
FSHBsts2	BV079389	GCCAGCTTCAGGCTAACATT	GACTTCATCTTGGGGTGGA	62	1101	G	d00621	FSHB	5	1102
P006A04sts1	BV079400	ATATCAGGTGCTCACAGTGC	GACTTAACTCTAGGAGTTCC	58	612	S	0893G05	BDNF	5	559
MYOD1sts3	BV012581	GGTGACTCAGACGCATCCA	ATAGGTGCCGTCGTAGCAGT	60	599	G	u12574	MYOD1	2	599
LDHAsts2	BV012579	TTTCACTGTCTAGGCTACAACAAGA	AGCTGGATAGTTGGCTGCAT	60	517	G	u07178	LDHA	1	517
RPS13sts1	BV079405	AGAGGCTGTGGATGACTCGT	GAAAGCATCTTGAAAGGAACAGA	62	902	L	Rattink et al, 2001	RPS13	12	839
P005E11sts1	BV079399	CTTCCCTAATGTCAGTG	ATTAAGAGACAGTAGAGTCC	50	799	S	0752C03	RPS13	1	630
BEQ0237A10	BV012569	GCTTGCATACCAATCTGCTG	ATGGCACTTTTGGTCTACCC	58	308	B	0237A10	sw747	0	308
BER0237A10	BV012571	AGCTTTGACTTATGCTGCTATGC	GAGCCAGATGCTGTGCTATG	58	400	B	0237A10	sw747	1	400
P010C10sts1	BV079412	GAGTACATTGACGCAATGGC	GAAAGCAACACATGATCGTG	58	920	S	0456F10	sw747	14	877
PTHsts1	BV079397	ACCAGGAAGAGATCTGTGAGTG	TGCCCTATGCTGTCTAGAGC	60	311	L	Lyons et al, 1997	PTH	5	308
AMPB3sts1	BV079377	CATTGTTGGAGCCAGGATCT	AGCCAATCAGAGGCTGAAAC	60	134	G	BG835121	AMPB3	0	134
ADMsts2	BV079374	ATTGAGAGACCGAGAGTCCG	TTGCTACTTCGCATATCACCC	58	646	G	d14875	ADM	4	645
GPX4sts1	BV012578	AGCTCAACAAGTGTGTGCTGA	GCCAAAGGGACCTTCCTC	60	146	L	Rattink et al, 2001	GPX4	0	146
DNASE2sts1	BV079387	CTCAGGGGCCAATTCAGACT	TTAGCAATCCTGAGGCAGGT	55	135	L	Rattink et al, 2001	DNASE2	0	135
CNN1sts1	BV079409	GGAGCACTACGAGGTCCAAC	CATGCAGTTTGCTCCCACT	60	198	L	Rattink et al, 2001	CNN1	0	168
PDE4Asts1	BV079396	GAAGTGGACATCCCATCACC	CCTCTTGATCGGTCTTCACC	60	202	L	Rattink et al, 2001	PDE4A	0	201
PRDX2sts1	BV079418	AGCTATGTCGCTCCAGGAAA	CCCTGTACTGACCCAGGAAA	60	140	L	Jorgenson et al, 1997	PRDX2	1	86

STS name	Accession number	Forward primer	Reverse primer	PCR temp. <sup>1</sup>	Product length (bp) <sup>2</sup>	Source Type <sup>3</sup>	Source Reference <sup>4</sup>	Locus <sup>5</sup>	# SNPs	of Contig length (bp) <sup>6</sup>
INSL3sts2	BV079410	CCCGTACTTCTCACCACCAT	ATCAGCCCATGGAAGAGATG	58	1130	G	x73636	INSL3	0	976
APBB1sts4	BV079372	GTGAGCAGTGGACACCGAGT	ATGATGAACGCAAAGGTGTG	60	271	G	BE235360	APBB1	2	271
CSF2sts1	BV079385	CAGCATGTGGATGCCATC	GTACAGCTTCAGGCGAGTCTG	60	973	L	Lyons et al, 1997	CSF2	10	972
IL4sts1	BV079417	GATCCCCAACCCTGGTTCTGCT	GGCAGAAAGACGTCGTCAC	62	433	L	Rettenberger et al., 1996	IL4	1	414
FOLR1sts1	BV012577	AGACGGTCCTTCTGCCTGT	TTGAGGAGGAGCCTATGGTTT	60	356	G	af054583	FOLR1	2	356
P006C12sts1	BV079398	TGAGTACTCGTTATGGACGC	CTGTGCCCTTTAGGACTGAGG	55	506	S	0185B06	FOLR1	18	460
P006D12sts1	BV079401	CCAAGATACAGAAGTAGGAGC	TGCAGTCTTCTTGGTGCAGG	55	393	S	0185B06	FOLR1	8	386
ADRB2sts1	BV079372	CAAGTACCAGAGCCTGCTGACC	TGAAGAAGGGCAGCCAGC	62	455	L	Lyons et al, 1997	ADRB2	0	454
CARSsts4	BV079411	CACTCGGAGGCCTACTTTGA	TGATTCTTCAGGCGTCTTT	55	134	G	BF077995	CARS	0	105
FKBP2sts3	BV079388	AAGAGGGACGGAGTTTGACA	TCTGATGGGATCACCAGCTT	62	263	G	BG609861	FKBP2	0	251
GSTP1sts1	BV079392	GGTTGTAGTCGGCAAAGGAG	ATGCCACCTCATCTACACC	55	157	G	BI184098	GSTP1	0	157
Hexbsts1	BV079415	CACTGGCACATAGTTGATGACC	TCCCTCGTAATCTGGCATATTC	55	<b>1000</b>	L	Lyons et al, 1997	Hexb	3	823

<sup>1</sup> Temperature in annealing step of PCR program.

<sup>2</sup> PCR product length in bp. Numbers in bold are estimated from agarose gel.

<sup>3</sup> Source type of the sequence that the primers were designed from: L: described in the literature; G: designed from Genbank entry; B: designed from BAC-end sequence and S: designed from sequence of subcloned BAC.

<sup>4</sup> Reference for the sequence that the primers were designed from. Lists the literature reference if source type is L; accession number if source type is G; or the BAC clone number for source types B and S. BAC clones belong to the library constructed by Rogel-Gaillard et al. (1999)

<sup>5</sup> The locus listed is used for placement on the map. It is either a locus from which the primers are designed or a locus that is present on the same BAC clone.

<sup>6</sup> Contig length reflects the length of the sequence contig containing the sequences that were compared to find SNPs.

## Results and discussion

### *PCR amplification and sequencing*

In total, 99 primer pairs were tested for amplification and subsequent sequencing of the PCR product. Sequence traces were obtained successfully for 67 PCR products (68%). All PCR products that were successfully sequenced are represented in table 1. For primer pairs originating from the literature or designed from Genbank entries or BAC-end sequences the success rate was around 80 percent. For primers designed from BAC subclone sequences the success rate of 51 % was considerably lower (Table 2). This lower rate is mainly due to the primer design strategy: most primers were designed to span the insert of the plasmid. The insert size of the plasmid, however, varied in the range from hundreds to several thousands of bases. Especially for the larger inserts no PCR product could be obtained. The 67 successfully amplified PCR products varied in length between 100 bp and 4 kb and had a total length of 43 kb.

### *Nucleotide variability*

For the eight animals in the SNP discovery panel all 67 PCR products were sequenced in both directions. This yielded 1072 reads of which 956 assembled in 73 distinct contigs. Six STSs were too long to be covered entirely by the sequence contigs. As a result, these STSs were analysed as two separate contigs, one starting from the forward primer site and one from the reverse primer site. Contigs varied in length between 86 bp and 1250 bp with an average of 444 bp and a total length of 32732 bp. In 48 STSs (72%) at least one polymorphism was identified. All PCR products over 500 bp in length contained at least one polymorphism, except for *INSL3sts2*, which contained no SNP in 976 bp of sequence. In the total contig length of 32429 bp 301 polymorphic positions were identified, an overall average of one SNP per 108 bp.

The four sequence data sources used for primer design can be divided in two categories: sequence data available in the public domain (both literature and Genbank) and sequence data generated at random in sample sequencing experiments (BAC-ends and subclones). The average SNP density revealed differences between the two data source categories: the STSs designed from public domain data yielded 1 SNP every 168 bp while the STSs designed from the random sample sequencing data yielded 1 SNP every 78 bp (Table 2).

The higher average SNP density in the STSs designed from sample sequencing data is mainly caused by a few STSs designed from subclones which are highly variable and contain up to 18 SNPs in only 460 bp (1 SNP every 25 bp). The number of highly variable STSs designed from data from the public domain is lower, probably due to the bias towards sequences related with functional properties, and thus with higher selection pressure. (Li and Sandler, 1991; Nickerson et al., 1998).

**Table 2:** Summary of the numbers of STSs with their success and polymorphism rates, grouped by Source type of the sequence that the primers were designed from.

Target region	Primer type <sup>1</sup>	Number of pairs tested	Number of pairs sequenced	of (%) Total product length <sup>2</sup>	Total contig length <sup>3</sup>	Number of SNPs	SNP density (bp/SNP) <sup>4</sup>	Poly-morphism rate <sup>5</sup>	
Public	L + G	56	45	(80)	16845	16845	100	168	559
Domain	L	31	25	(80)	7911	7911	56	141	469
	G	25	20	(80)	8934	8934	44	203	647
Sample	B + S	43	22	(51)	15584	15584	201	78	262
Sequencing	B	5	4	(80)	1262	1262	2	631	2094
	S	38	18	(47)	14322	14322	199	72	244
Total	L + G + B + S	99	67	(67)	32429	32429	301	108	361

<sup>1</sup> L: described in the literature; G: designed from Genbank entry; B: designed from BAC-end sequence and S: designed from sequence of subcloned BAC.

<sup>2</sup> Total PCR product length in bp.

<sup>3</sup> Total contig length of the sequence contig analysed.

<sup>4</sup> Average SNP density.

<sup>5</sup> Average SNP density normalised for sample size.

Fahrenkrug et al. (2002) reported a polymorphism rate of one SNP in every 184 bp of EST-based PCR products in pigs. This intragenic polymorphism rate (1/184) is highly comparable to the SNP density in the STSs designed from public domain data in this study (1/168). A direct comparison between the results of these studies, however, is only valid because both studies used a panel of eight diploid individuals for SNP discovery. Because the number of polymorphisms to be observed is strongly dependent on the number of chromosomes sampled these numbers should be normalised for the assayed sample size. Nei and Li (1979) introduced an index termed nucleotide diversity or heterozygosity. This value reflects the rate of nucleotide difference between two randomly chosen chromosomes. The overall polymorphism rate of 1 SNP every 108 bp as found in this study relates to a heterozygosity value of 1/361 (1 SNP every 357 bp). For the STSs designed from public domain data the heterozygosity is 1/559. The numbers reported by Fahrenkrug et al. (2002) translate in a heterozygosity value 1/609. For other mammals similar values have been reported: in mice it is 1/515 (Lindblad-Toh et al., 2000), in cattle 1/443 (Heaton et al., 2001) and in humans 1/1331 (Sachidanandan et al., 2001). The lower heterozygosity in human as compared to other species might reflect the reduced diversity in humans, probably due to recent population expansion (Kaessmann et al., 2001). Composition of the polymorphisms identified in this study was 63% C/T or A/G, 13% G/T or A/C, 7% G/C, 7% A/T and 10% indels. One polymorphism showed more than two alleles. This occurrence of the different allelic combinations is highly comparable with the data as described by Fahrenkrug et al. (2002).



### *Clustering of the SNP*

Although at some positions polymorphisms tend to cluster, every polymorphic position was considered as a separate SNP. In this study 16 clusters of two directly neighbouring SNPs were identified, clusters of three SNPs were found five times and there were 2 clusters of 4 SNPs. The remaining 246 SNPs do not have a SNP on the neighbouring positions.

For the application of SNPs in genotyping assays like allele specific amplification (ASO; Newton et al., 1989) or Single Base extension (SBE; Syvänen et al., 1990) it is important that the SNPs are not too close together because for both methods a primer has to be designed immediately adjacent to the SNP. The presence of another SNP within approximately 20 bp will limit the possibilities for designing the genotyping primer. In this study, 115 SNPs do not have a SNP within 20 bp on either side, thus leaving two possibilities for primer design. For an additional 138 another polymorphism is present at one side, leaving the other side as an option for primer design. As a result, 253 of the SNPs identified offer the opportunity to design a SBE primer. A list of all SNP and their flanking sequence is available from Table S1.

### *Polymorphism reliability*

Although regions of lower quality were masked some of the observed sequence differences might reflect sequence artefacts and might not be true SNPs. For 198 SNPs, however, the panel represented a sample that is homozygous for the minor allele. The occurrence of homozygotes for both the major and the minor allele strongly suggests that most polymorphisms identified in this study are SNPs rather than a sequence artefact. For several positions this suggestion is even strengthened by the presence of one or more heterozygous samples.

For 20 SNPs, SBE assays were developed using the SNaPshot kit (ABI, Foster City, CA, USA) in a multiplex format as described by Jungerius et al. (2003). For 18 SBE primers the revealed genotypes were confirmed by re-sequencing of the assayed samples (data not shown). The alleles assigned to several samples from a family structure followed Mendelian inheritance. Two primers failed to yield clear genotypes: one primer failed to yield a product and one primer only yielded ambiguous signals.

### *Relevance of the SNPs*

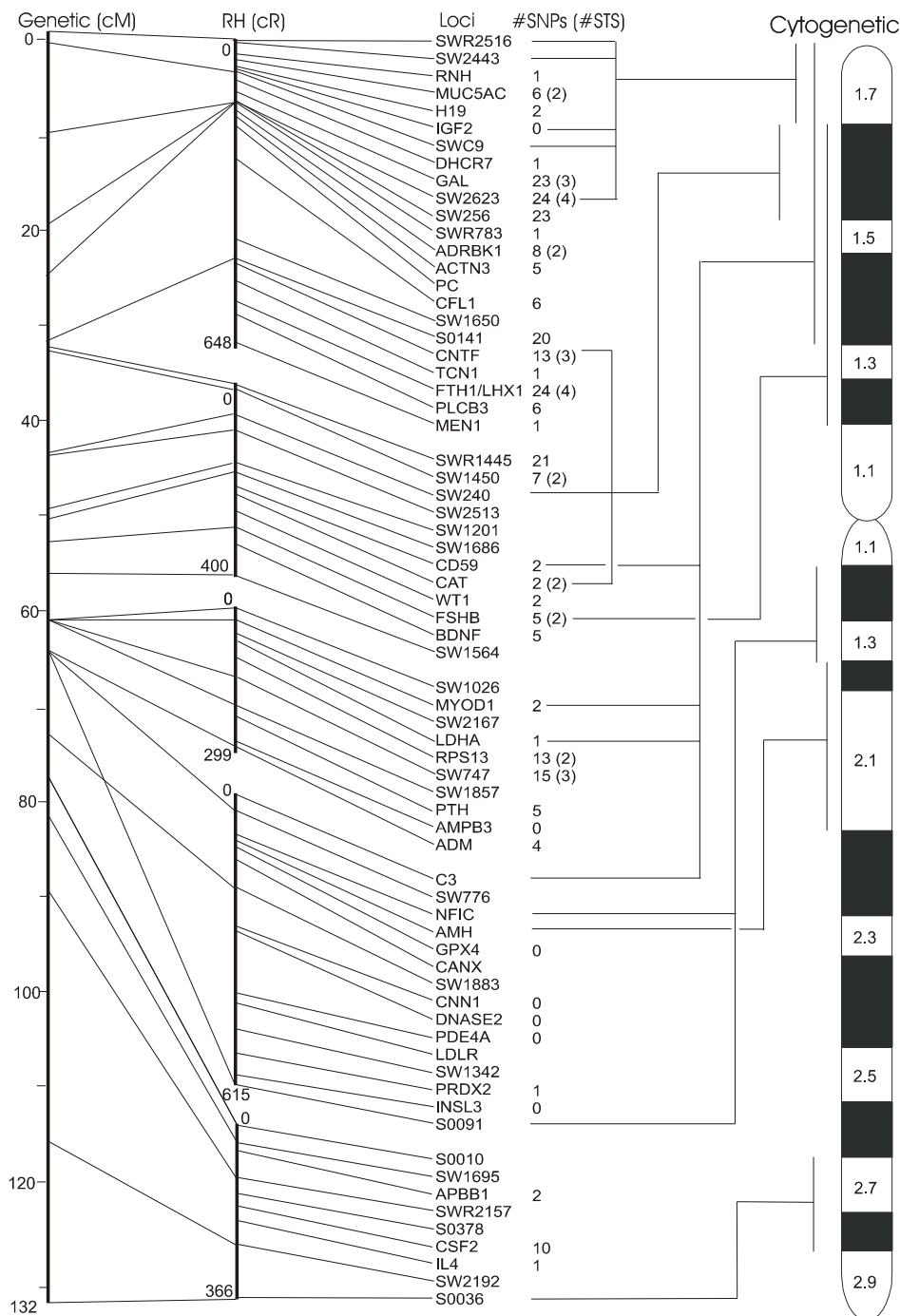
In the SNP discovery panel five different Dutch breeding companies each were represented by a Large White. The other three, Meishan, Pietrain and Wild Boar, are more distantly related and therefore a higher polymorphism rate is expected. For all eight animals, the genotypes identified are represented in Table S1.

For 191 SNPs (63%) both the major and the minor allele were found within the five Large Whites. Because both alleles are present within commercial White lines these

SNPs are particularly relevant for the genetic analysis of commercial breeds. For 269 SNPs (89%) both the major and the minor allele were found within the group of the Meishan plus the five Large Whites. For 207 SNPs (69%) both the major and the minor allele were found within the group of the Pietrain plus the five Whites. For 210 SNPs (70%) both the major and the minor allele were found within the group of the Wild Boar plus the five Whites. Because many of the experimental populations to identify QTL are based on crosses between these breeds, these SNPs are highly relevant for the fine mapping of the QTL identified within these crosses.

### *Genomic position of the SNPs*

Primer pairs used in this study were designed from sequence data originating from SSC2 or were designed from genes likely to map to SSC2 based on the human-porcine comparative map. All PCR products were either directly mapped on a Radiation Hybrid (RH) panel (Rattink et al., 2001b) or were mapped indirectly as they were derived from a BAC containing a locus that was previously mapped on the RH panel. For 12 loci more than one PCR product was used. Together, the 67 PCR products originate from 47 distinct chromosomal locations. An overview of the distribution of the locations over the chromosome and the number of identified SNPs is shown in Figure 1.



**Figure 1:** Map of SSC2 giving an overview of the distribution of loci over the chromosome. Loci names followed by numbers represent loci investigated in this study. The numbers indicate the number of SNPs identified at this locus. The number of STSs per locus is indicated between brackets, if not equal to one. STSs not mapped to SSC2: FOLR1sts1 (3 SNPs), P006C12sts1 (8), P006D12sts (18) and APBB1sts4 (4) map to SSC9. ADRB2sts1 (0), CARSts4 (0), FKBP2sts1 (0), GSTP1sts1 (0) and HEXBsts1 (4) have not been mapped.

In total, nine STSs were not mapped to SSC2. Five STSs (ADRB2sts1, CARSts4, FKBP2sts3, GSTP1sts1, and HEXBsts1) could not be mapped on the Radiation Hybrid Panel because of coamplification of the endogenous gene from the hamster host cells. Some of the loci were included in this study based on the comparative map between pig and human. This comparative map, however, shows that a breakpoint between two groups with conserved gene order is present between *GAL* and *FOLR1*. Although *GAL* and *FOLR1* are in the same region on HSA11, *GAL* maps to SSC2 (Figure 1) but *FOLR1* maps to SSC9. For *FOLR1* three STSs were sequenced: FOLR1sts1, P006C12sts1 and P006D12sts1, with 3, 8 and 18 SNPs respectively. On HSA11 *APBB1* is located between *RNH* and *ADM*, which both map to SSC2. *APBB1*, however, maps to SSC9 due to a rearrangement of a small fragment containing *APBB1*. For *APBB1* one STS was sequenced: APBB1sts4, which contained 4 SNPs. In conclusion, over 300 SNP loci were identified through comparative re-sequencing of 67 PCR products. 268 SNP loci occur as clusters corresponding to the screened PCR fragments originating from 41 distinct regions on SSC2. 29 SNP loci occur as clusters from two distinct regions on SSC9. The identified SNPs might be relevant for linkage and association studies in both commercial populations and experimental test crosses.

### Acknowledgements

We thank Barbara Harlizius and Tineke Veenendaal for their useful comments and technical assistance.

### Supplementary material

The following material is available from <http://www.blackwellpublishing.com/products/journals/suppmat/AGE/AGE1056/AGE1056sm.htm>

Table S1 List of all identified polymorphisms with their flanking sequences and an overview of the alleles found in the sequenced individuals.

### References

- Archibald AL, Haley CS, Brown JF, Couperwhite S, McQueen HA, Nicholson D, Coppieters W, Van de Weghe A, Stratil A, Wintoro AK, et al. (1995) The PiGMaP consortium linkage map of the pig (*Sus scrofa*). *Mammalian Genome* **6**, 157-75
- Bonfield JK, Smith KF and Staden R (1995) A new DNA sequence assembly program. *Nucleic Acids Research* **24**, 4992-99
- Bonfield JK and Staden R (1996) Experiment files and their application during large-scale sequencing projects. *DNA Sequence* **6**, 109-17.

- De Koning DJ, Janss LL, Rattink AP, van Oers PA, de Vries BJ, Groenen MA, van der Poel JJ, de Groot PN, Brascamp EW, van Arendonk JA. (1999) Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (*Sus scrofa*), *Genetics* **152**, 1679-90
- Ewing B, Green P. (1998a) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**(3),186-94.
- Ewing B, Hillier L, Wendl MC, Green P. (1998b) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**(3), 175-85.
- Fahrenkrug SC, Freking BA, Smith TP, Rohrer GA, Keele JW. (2002) Single nucleotide polymorphism (SNP) discovery in porcine expressed genes. *Animal Genetics* **33**, 186-95
- Hawken RJ, Murtaugh J, Flickinger GH, Yerle M, Robic A, Milan D, Gellin J, Beattie CW, Schook LB, Alexander LJ. (1999) A first-generation porcine whole-genome radiation hybrid map. *Mammalian Genome* **10**, 824-30
- Heaton MP, Grosse WM, Kappes SM, Keele JW, Chitko-McKown CG, Cundiff LV, Braun A, Little DP, Laegreid WW. (2001) Estimation of DNA sequence diversity in bovine cytokine genes. *Mammalian Genome* **12**(1), 32-7
- Jorgensen CB, Wintero AK, Yerle M, Fredholm M. (1997) Mapping of 22 expressed sequence tags isolated from a porcine small intestine cDNA library. *Mammalian Genome* **8**, 423-7
- Jungerius BJ, Veenendaal A, van Oost BA, te Pas MFW, Groenen MAM. (2003) Typing single nucleotide polymorphisms using a gel-based sequencer: a new data analysis tool and suggestions for improved efficiency. *Molecular Biotechnology* **25**, 283-287
- Kaessmann H, Wiebe V, Weiss G, Paabo S. (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genetics* **27**(2), 155-6
- Li W.H. and Sadler L.A. (1991) Low nucleotide diversity in man. *Genetics* **129**(2), 513-23
- Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette JP, Ardlie K, Reich DE, Robinson E, Sklar P, Shah N, Thomas D, Fan JB, Gingeras T, Warrington J, Patil N, Hudson TJ, Lander ES. (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics* **24**(4), 381-6
- Lyons LA, Laughlin TF, Copeland NG, Jenkins NA, Womack JE, O'Brien SJ. (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genetics* **1**, 47-56
- Moran C (1993) Microsatellite repeats in pig (*Sus domestica*) and chicken (*Gallus domesticus*) genomes. *Journal of Heredity* **84**, 274-80

Nei, M and Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**(10) 5269-73.

Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, Markham AF (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Research* **17**(7):2503-16

Nezer C, Moreau L, Brouwers B, Coppieters W, Dettleux J, Hanset R, Karim L, Kvasz A, Leroy P, Georges M. (1999) An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nature genetics* **21** ,155-6

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genetics* **19**(3), 233-40.

Rattink AP, De Koning DJ, Faivre M, Harlizius B, van Arendonk JA, Groenen MA. (2000) Fine mapping and imprinting analysis for fatness trait QTLs in pigs. *Mammalian Genome* **11**, 656-61

Rattink AP, Jungerius BJ, Faivre M, Chardon P, Harlizius B, Groenen MA (2001a) Improving the comparative map of SSC2p-q13 by sample sequencing of BAC clones. *Animal Genetics*. **32**(5):274-80

Rattink AP, Faivre M, Jungerius BJ, Groenen MA, Harlizius B. (2001b) A high-resolution comparative RH map of porcine chromosome (SSC) 2. *Mammalian Genome* **12**, 366-70

Rettenberger G, Bruch J, Fries R, Archibald AL, Hameister H. (1996) Assignment of 19 porcine type I loci by somatic cell hybrid analysis detects new regions of conserved synteny between human and pig. *Mammalian Genome* **7**, 275-9

Rieder MJ, Taylor SL, Tobe VO, Nickerson DA. (1998) Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Research* **26**(4), 967-73

Rogel-Gaillard C, Bourgeaux N, Billault A, Vaiman M, Chardon P (1999) Construction of a swine BAC library: application to the characterization and mapping of porcine type C endoviral elements. *Cytogenetics and Cell genetics* **85**, 205-11

Rohrer GA, Alexander LJ, Hu Z, Smith TP, Keele JW, Beattie CW. (1996) A comprehensive map of the porcine genome. *Genome Research* **6**(5):371-91

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley

DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D; The International SNP Map Working Group. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**(6822), 928-33

Syvanen AC, Aalto-Setälä K, Harju L, Kontula K, Soderlund H (1990) A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics*. **8**(4):684-92

Winterø AK, Fredholm M, Thomsen PD. (1992) Variable (dG-dT)<sub>n</sub>.(dC-dA)<sub>n</sub> sequences in the porcine genome. *Genomics* **12**(2), 281-8

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES, et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**(5366), 1077-82.





## **Chapter 3**

### **POSA:**

## **Perl Objects for DNA Sequencing Data Analysis**

Jan Aerts, Bart Jungerius, Martien Groenen

Animal Breeding and Genetics Group, Wageningen University

PO box 338, NL-6700 AH Wageningen, Netherlands

BMC Genomics 5: 60 (2004)

### Abstract

Capillary DNA sequencing machines allow the generation of vast amounts of data with little hands-on time. With this expansion of data generation, there is also a growing need for automated data processing. Most available software solutions, however, still require user intervention or provide modules that need advanced informatics skills to allow implementation in pipelines.

Here we present POSA, a pair of new perl objects that describe DNA sequence traces and phrap contig assemblies in detail. These objects allow a flexible and easy design, implementation and usage of perl-based pipelines, while requiring only little programming skills. Methods included in POSA include basecalling with quality scores (by *Phred*), contig assembly (by *Phrap*), generation of primer3 input and automated SNP annotation (by *PolyPhred*).

## Background

Today, many genetics laboratories have access to modern capillary DNA sequencing machines, such as the ABI PRISM 3100, 3700 or 3730. These machines generate vast amounts of raw sequence data with little user intervention. Consequently, the amount of data to be analysed has expanded and the bottleneck now is the analysis capacity.

Data analysis capacity can be increased by higher levels of automation. Investments in infrastructures to process the raw sequencing data in sophisticated but rigid pipelines might be justified for larger laboratories and larger projects but might be too costly for smaller laboratories. In addition, rigid pipelines are too impractical for the variety of different projects that share run-time on the same machine but require (slightly) different analysis procedures (e.g. vector trimming is needed in plasmid sequencing, but needless when sequencing PCR products).

Nucleotide sequence analysis can be performed with a variety of software tools. Although the number of console and web-based software tools has grown rapidly, the routine use of data input, output and storage may be inconvenient. Furthermore, for performing a series of analyses with different software tools, the sequence data need to be reformatted to the required data structure. Alternatively, sophisticated software suites that do provide an integrated environment often are expensive.

Several of the available software solutions are designed to facilitate automated DNA sequence analysis at low cost. Well-known solutions are the Staden package and bioperl.

The Staden Package contains *pregap4* and *gap4*, full-featured applications with an intuitive graphical user interface (Staden *et al.*, 2000). These programs handle a list of raw sequence reads method-by-method. The programs in the Staden Package typically require a degree of user intervention and thus hands-on time.

Alternatively, Bioperl is a group of perl modules describing many genetics and genomics concepts (Stajich *et al.*, 2002). For example, it includes the `Bio::Seq::SeqWithQuality` object that provides some of the basic properties of a raw sequence (i.e. its nucleotide sequence and quality values). The available modules, however, only offer a modest selection of basic sequence properties and require a certain level of informatics skills.

Smaller laboratory sites, however, often need to implement versatile pipelines that can be adjusted for any research question that suits the project best. Most available software solutions do not comply with these requirements: a highly automated sequence analysis pipeline that describes a raw sequence and sequence contig in detail, but still is flexible in design and relatively simple to use.

Here, we present POSA, a set of two new perl objects (`Read.pm` and `Contig.pm`) that describe a raw sequence and a phrap contig in detail and are easily implemented in perl-based pipelines.

## Implementation

The POSA source code is entirely coded in object-oriented Perl and consists of two objects: `Read.pm` and `Contig.pm`. In general, there are two important concepts associated with objects: methods (built-in procedures that can be performed on the object) and properties (describing some of the characteristics of the object). Most methods in the objects rely on the availability of other third-party programs (see Dependencies). Basically, POSA provides a wrapper around these programs and provides easy design and implementation of these programs in automated data analysis. The `Read.pm` object describes a DNA sequence trace and includes methods for data import from in a variety of formats. It relies on *Phred* (Ewing *et al.*, 1998a and 1998b) for import and interpretation of raw sequence data. The original trace data are stored in binary (*scf*) format within the object. Other methods of `Read.pm` use modules of the Staden Package (Staden *et al.*, 2000), such as *qclip* and *vector\_clip* (if installed). Properties of `Read.pm` include e.g. the DNA sequence, quality scores, template and vector names and read direction.

The `Contig.pm` object contains a method to assemble contigs of reads using the *Phrap* program (Ewing and Green, 1998). The object typically is created based on a list of `Read.pm` objects and can be exported as alignments or screened for polymorphisms using *PolyPhred* (Nickerson *et al.*, 1998).

Both the `Read.pm` and `Contig.pm` objects were designed with flexibility in mind. To allow a (virtually) unlimited amount of data to be processed, the perl scripts using these objects will work sequence-by-sequence rather than method-by-method. Typically, these objects are called from straightforward perl scripts that outline the analysis steps to be performed. Example scripts using the objects can be accessed from the download website. An example of a script using the two objects to process a set of reads and annotate sequence polymorphisms from the assembled contig is given in Table 1.

POSA was developed with perl 5.6.1 and tested on a SuSE linux 8.1 system for *abi*-files from the ABI PRISM 377 DNA Sequencer and 3100 Genetic Analyzer (Applied Biosystems). *Phred*, *Phrap* and *PolyPhred* versions were 0.000925.c, 0.990329 and 4.05, respectively.

## Results and Discussion

### *Functionality*

POSA provides an interface to design and implement automated sequence data analysis. Sequence data may be used in a variety of formats and be from a variety of

sources, e.g. data in *fasta*, *abi/ab1* or *scf* format retrieved from websites or of newly generated traces. In addition, new objects can be initiated from a text file or can be opened from previous stored objects. Subsequently, a variety of methods can be applied, including basecalling and assessment of quality codes (by *Phred*), quality clipping, vector clipping, screening for *E. coli* (or other) sequence, contig assembly (by *Phrap*) and analysis. The method *asPrimer3* can automatically generate input for the primer3 program (Rozen and Skaletsky, 2000) and is available in both objects.

To facilitate automated SNP discovery or typing, the *SearchSnps* method will generate output as shown in Table 1. This method is based on the *PolyPhred* program and uses the 'rank' argument to set the stringency.

Finally, read data can be stored in objects, or in files in either *exp*, *scf* or *fasta* format. In addition, the data can be saved in a *primer3* input file to allow automated PCR primer design, or data can be saved in MIPE format (i.e. an XML format to store information on PCR experiments; see <http://mipe.sourceforge.net>). Data on assembled contigs can be exported as a list of reads in a contig, as consensus sequence, as alignment, as putative SNPs, as SBE primers for SNP genotyping, as *gff* file for visualization in Gbrowse (Stein et al., 2002). Combinations of the diversity of input, analysis and output options allow for a wide spectrum of possible implementations. Examples of possible analysis pipelines include (but are not limited to) BAC-end sequencing with automated PCR primer design for BAC contig building or for chromosome walking and resequencing of PCR products with SNP annotation either for SNP genotyping or for SNP discovery and SBE primer design. For these examples, perl scripts are provided on the web site (<http://posa.sourceforge.net>).

### *Performance*

Although it represents only one of the numerous possible POSA-based pipelines, performance of POSA was validated by comparison of SNP discovery with the data after analysis using the Staden package. To do so, 5 PCR products were resequenced from a panel of 16 individuals to identify SNPs. Manual editing using the Staden Package revealed a total of 48 SNPs. Automated analysis using POSA also yielded a total of 48 SNPs with SNP ranking codes 1-3. Together, 41 SNPs were assigned with both manual editing and POSA. The remaining 7 SNPs assigned in manual editing corresponded to SNPs with ranks 4-6 in the POSA analysis. The 7 SNPs that were only assigned by POSA all originated from regions with lower quality sequence. While analysis time was reduced from several hours to few minutes, POSA assigned SNPs in a way that was highly consistent with manual editing. This was expected because POSA provides options for an integrated analysis pipeline, but basically is a wrapper around well-established sequence analysis tools like *Phred*, *Phrap* and *PolyPhred*.

**Table 1:** A typical script (upper cell) that takes a list of ab1 files for analysis and assembly, and reports the contig, a lists the putative SNP positions and SBE primers (lower cell).

```
#!/usr/bin/perl
use strict;
use warnings;
#Use the POSA perl objects
use POSA::Read;
use POSA::Contig;
#What is the file containing the names of all abi-files?
my $file_of_filenames = shift;
#Foreach read: create a POSA::Read object and add it to a @reads array.
my @reads = ();
open FOFN, $file_of_filenames;
foreach my $abi_file ( <FOFN> ) {
    chomp $abi_file;
    my $read = POSA::Read->importFromAbi(name => $abi_file, file => $abi_file);
    push @reads, $read;
}
close FOFN;
#Create a list of POSA::Contig objects based on all POSA::Read objects
my @contigs = POSA::Contig->phrap(reads => \@reads, gap_init_penalty => -6);
#Foreach contig: search for SNPs and produce some output
foreach my $ctg ( @contigs ) {
    #Run the command to annotate SNPs in this contig
    $ctg->searchSnps(rank => 6);
    #Print output
    print $ctg->asGenotypes, "\n";
    print $ctg->asSbe(length => 35), "\n";
    print $ctg->asAlignment, "\n";
}

#####GENOTYPES#####
Contig1
Polymorphisms: 25(1),133(3),292(2)
Genotypes:      25      133      292
sample5.ab1     G/G      T/T      G/G
sample6.ab1     G/G      G/T      A/A
sample10.ab1    G/G      T/T      G/G
sample11.ab1    C/G      T/T      G/G
#####POSSIBLE SBE PRIMERS#####
25      FORWARD      CCCTCTGCAat
25      REVERSE      GCAGCAGGAAGAGGCAGGGCAGTGCCACGGGCTCC
133     FORWARD      aCCCCCCCCGCGTCAAATGG*AGCAAGGTGCGCTC
133     REVERSE      CAGGATGGGGACGTCCT*CCC*TCTGCCCCGCTGGC
292     FORWARD      ACGCTGCTGTGTCGCGCCGCCGCCAGCGATGC
292     REVERSE      TCGATGCCAGCCACCACCTCGCAGCGGTACAGCCC
#####ALIGNMENT#####
sample5.ab1      ccagcagtgcg*aGTGGGGCTGGGGGAGCCCGTGGCACTGCCCTGCCT-->
sample6.ab1      cCAGgcagTGCG*AGTGGGGCTGGGGGAGCCCGTGGCACTGCCCTGCCT-->
sample10.ab1                      agTGGGGCTGGGGGAGCCCGTGGCACTGCCCTGCCT-->
sample11.ab1                      tgGGGCTggcgggaGCCCGTGGCACTGCCCTGCCT-->
Contig1          cCAGgcaGTGCG*AGTGGGGCTGGGGGAGCCCGTGGCACTGCCCTGCCT-->
```

### *Intended use and benefits for users*

POSA is a tool that provides easy and highly automated DNA sequence and contig data analysis using popular analysis tools. Automated sequence analysis reduces analysis time from several hours to a few minutes. Design of the pipeline can easily be expanded or adapted through perl scripts. No enhanced programming skills are required as the perl scripts are relatively easy to use and modify, while it relies on well-established software modules (eg. *Phred* and *Phrap*). Overall, this guarantees easy implementation of highly automated quality pipelines in combination with high flexibility in setup and design.

The perl objects are available under an open source license allowing code improvements by the user community.

### **Conclusions**

POSA describes a DNA sequence read and a *phrap* contig assembly in detail. These objects allow a flexible and easy setup of perl-based pipelines including generating primer3 input and automated SNP discovery, while requiring only little programming skills.

### **Availability and requirements**

Project name: POSA  
Project home page: <http://posa.sourceforge.net>  
Operating system: platform independent  
Programming language: Perl 5.6.1  
License: Artistic License

#### Requirements:

Perl modules: Carp; Statistics::Descriptive; Tie::File; IO::File; POSIX::Storable.  
Phred, Phrap, PolyPhred  
Pregap4 , gap4 (Staden Package (optional))  
Primer3 (optional)

### **List of abbreviations**

POSA Perl objects for sequence analysis  
SNP single nucleotide polymorphism  
abi/ab1 ABI PRISM trace file format  
scf standard chromatogram format  
exp experiment file format, developed by Staden (see <http://staden.sourceforge.net>)  
MIPE minimum information on PCR experiments (see <http://mipe.sourceforge.net>)  
BAC bacterial artificial chromosome  
PCR polymerase chain reaction  
SBE single base extension

### **Authors' contributions**

JA programmed the Perl objects and participated in development of concept and architecture of the software; BJ participated in development of concept and architecture and wrote the manuscript; MG supervised the project. All authors read and approved the final manuscript.

### **Acknowledgements**

The authors wish to thank Tineke Veenendaal for testing.

### **References**

- Ewing B, Hillier L, Wendl MC, Green P. (1998a) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8(3):175-85.
- Ewing B, Green P. (1998b) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8(3):186-94.
- Nickerson DA, Tobe VO, Taylor SL. (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* 25(14):2745-51.
- Rozen S, Skaletsky H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol.* 132:365-86.
- Staden R, Beal KF, Bonfield JK. (2000) The Staden package, 1998. *Methods Mol Biol.* 132:115-30.



Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.* 12(10):1599-610.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12(10):1611-8.



## **Chapter 4**

### **Typing Single Nucleotide Polymorphisms using a gel-based sequencer: a new data analysis tool and suggestions for improved efficiency**

Bart Jungerius<sup>1</sup>, Tineke Veenendaal<sup>1</sup>, Bernard van Oost<sup>2</sup>,  
Marinus te Pas<sup>3</sup>, Martien Groenen<sup>1</sup>

<sup>1</sup> Department of Animal Sciences, Wageningen University, Wageningen, The Netherlands

<sup>2</sup> Faculty of Veterinary Medicine, Utrecht University, The Netherlands

<sup>3</sup> ID-Lelystad, Institute for Animal Science and Health, Lelystad, the Netherlands

Molecular Biotechnology, 25: 283-287 (2003)

### **Abstract**

Single Nucleotide Polymorphisms (SNPs) are increasingly used as genetic markers. Although a high number of SNP genotyping techniques have been described, most techniques still have low throughput or require major investments. For laboratories that have access to an automated sequencer, a Single Base Extension (SBE) assay can be implemented using the ABI SNaPshot kit. Here we present a modified protocol comprises of multiplex template generation, multiplex SBE reaction and multiplex sample analysis on a gel-based sequencer like the ABI 377. These sequencers run on a Macintosh platform, but on this platform the software available for analysis of data from the ABI 377 has limitations. First, analysis of the size standard included with the kit is not facilitated. Therefore a new size standard was designed. Second, using Genotyper (ABI) the analysis of the data is very tedious and time-consuming. To enable automated batch analysis of 96 samples, with 10 SNPs each, we developed SNPtyper. This is a spreadsheet-based tool that uses the data from Genotyper and offers the user a convenient interface to set parameters required for correct allele calling. In conclusion, the method described will enable any lab having access to an ABI sequencer to genotype up to 1000 SNPs per day for a single experimenter, without investing in new equipment.

Single Nucleotide Polymorphisms (SNPs) are increasingly used as genetic markers in a wide variety of species. An efficient genotyping system is required to employ SNP markers in genetic studies, which requires the genotyping of many SNPs in large numbers of individuals. Several dozens of SNP genotyping techniques have been described over the last few years. The traditional method for SNP genotyping is PCR-RFLP which provides relatively robust signals through a simple method with low costs. This method, however, has limited throughput, is more difficult to automate and not all SNPs disrupt or create a restriction site. (1). Hybridisation-based assays are widely used in a variety of assays (2-7), but compared to enzyme-based methods these are less robust and have a less specific allelic distinction. (8). The single base extension (SBE) method, also called minisequencing, relies upon DNA polymerase and is adapted to various detection platforms ranging from a relatively low- throughput ELISA-like assay (9), to more expensive high-throughput mass spectrometry (10). SNP genotyping methods have been reviewed extensively (11,12) and in conclusion most methods still have limited throughput capacity or require major investments in new and often expensive equipment. Especially for smaller laboratories the usage of existing machinery for implementation of SNP genotyping is favourable above investing in new techniques, because no method described so far seems to be the sole promising method for future high-throughput SNP genotyping (12). For laboratories that have access to an automated sequencer, the Single Base Extension (SBE) assay is an attractive SNP genotyping method. Because of its universal conditions and its high specificity, this method efficiently facilitates multiplexing (13). The individual products of this multiplex reaction can be detected on a fluorescent sequencer when a mixture of primers, each with a different length, is extended with fluorescently labelled dideoxynucleotides (ddNTPs) (14).

A convenient way to implement the SBE assay is by applying the ABI Prism® SNaPshot™ chemistry (Applied Biosystems, Foster City, CA). Application of this kit on gel-based fluorescent sequencers, however, has some limitations in data analysis. To overcome these limitations and to increase genotyping throughput, the SNaPshot protocol was adapted for multiplex SNP genotyping on an ABI377. In addition, a batch process for data analysis was designed that enabled flexible and robust genotype assignment requiring low hands-on time.

To increase genotyping throughput a multiplex approach was developed for both the SBE reaction and the amplification of the template DNA for the SBE reaction.

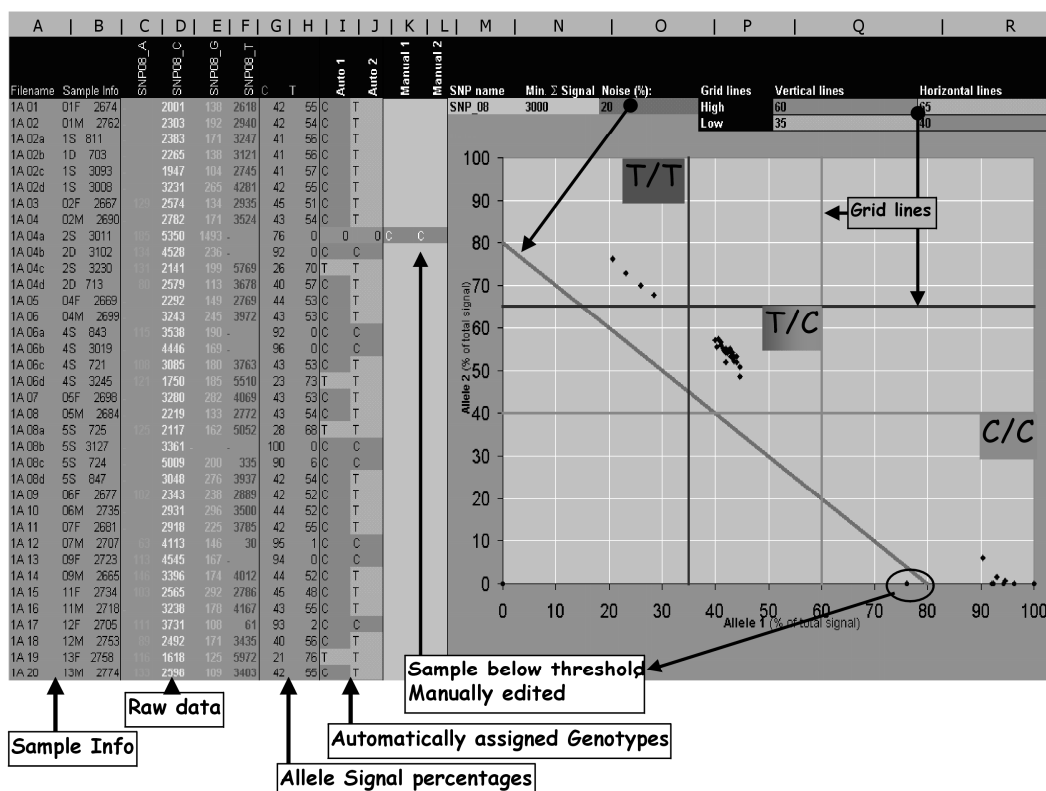
Amplification of the templates was carried out in two separate multiplex PCR reactions, which together represented up to 10 sequence tagged sites (STSs) that harboured the SNPs under investigation. Depending on the product sizes either Accuprime Supermix I (<200 bp) or Accuprime Supemix II (>200 bp) (Invitrogen, Carlsbad, CA, USA) was used according to the manufacturer's instructions. Using this kit, optimisation of the multiplex

reactions merely consisted of subtle changes in relative primer concentrations. A more detailed protocol for the development of a multiplex PCR assay has been described by Henegariu et al. (15). After amplification, portions of both reactions were pooled. Excess primers and dNTPs were enzymatically removed using Exonuclease I (ExoI; USB, Cleveland, OH, USA) and shrimp alkaline phosphatase (SAP; USB, Cleveland, OH, USA).

The initial version of the SNaPshot kit was designed to analyse a single SNP in a single tube. Multiplex SBE reactions assaying 4 SNPs have been reported previously (16). The number of SNPs per reaction tube, however, was successfully increased to 10 SNPs, even with the initial singleplex kit. For multiplexing, SBE primers had a locus specific 3'-tail of 19 to 24 bases in length and all had a melting temperature between 50 and 55 °C. To facilitate gel-based separation of the different reaction products the SBE primers had non-specific 5'-tails to give them a total length of 25 to 70 bases in length with 5 base differences. Performing the SNaPshot reaction in a multiplex format both increases capacity and decreases reaction cost per genotype. A further decrease in reaction costs was achieved by performing the ten-plex SBE reactions with three quarters of the SNaPshot ready reaction mix replaced by the cheaper Half-BD Buffer (Genetix, New Milton, UK). This replacement did not influence signal intensity or accuracy. After the reaction excess ddNTPs were enzymatically removed using SAP.

Reaction products were run in a 48-well format for 1.5 hours on a 36 cm 4.75% denaturing Long Ranger gel (BMA, Rockland, ME, USA) on an ABI Prism 377 automated sequencer (Applied Biosystems). Before samples were loaded on gel an internal size standard was added. Although a size standard is included in newer versions of the SNaPshot kit, a custom-made size standard consisting of four TAMRA-labelled oligonucleotides with lengths of 15, 20, 75 or 80 bases was used. This was necessary because the SNaPshot size standard applies ABI's fifth dye chemistry but this fifth dye colour (and therefore the size standard) can not be analysed with the latest version of Genescan for Macintosh (version 3.1.4; ABI, Foster City, CA, USA).

After sample file extraction and product length assignment using Genescan, the trace files are imported into Genotyper (version 2.0; ABI). This program was set to analyse ten consecutive 5-bp windows (from 25-30 bp to 70-75bp). For each of the four colours, it identifies and labels the largest signal peak in each of the ten windows. For each colour-window combination both the product size and the peak area are represented in a row of the output table. This table will therefore consist of 40 rows per sample, each containing data on a single peak.



**Figure 1:** SNPtyper screenshot of a typical SNP analysis worksheet. Each line represents a sample. Columns A and B contain sample information; columns C through F show the raw data of the signal values. In columns G and H the headers contain the two alleles under consideration (here C and T) and the remainder shows the percentile values for these alleles. Columns I and J express the automated allele calling. In the scatter plot the X-axis shows the signal value of the first allele (C) while the Y-axis shows the signal for the second allele (T). The remaining five criteria were set accordingly: noise allowance: 20% (represented by the diagonal line) and four values corresponding to four coloured lines that form a grid: 60, 35, 65 and 40 %. The grid formed by four coloured lines divides the plot area in 9 boxes. Samples plotted in the upper-left box are assigned homozygous for allele 2 (T/T). Samples plotted in the centre box are assigned heterozygous (C/T). Samples plotted in the lower-right box are assigned homozygous for allele 1 (C/C). The noise of the sample shown below the diagonal green line exceeded the noise allowance threshold and the alleles were assigned manually.

For batch analysis of this extensive Genotyper output table we designed SNPtyper, a spreadsheet-based template that facilitates simultaneous analysis of up to 96 samples. In brief, the Genotyper data table is imported in the first SNPtyper worksheet. From here, all data from one sample were regrouped automatically on a single row and were redistributed using a separate worksheet for each SNP. A screenshot of a typical SNPtyper worksheet is shown in Figure 1. On these SNP worksheets the variables and parameters required for the analysis of each SNP marker were selected. First, the two

alleles to be considered were selected from a pull-down menu. Next, the minimum value for the total sum of the four signal values is entered (Min. S signals). The sum of all four signals is the only variable reflected by an absolute value; all other variables are expressed as a percentage of this value. All samples with signals not exceeding the minimum value threshold are low signal data and are excluded from further analysis. Remaining samples are represented in a scatter plot. In this plot, the X-axis represents the signal value for the first allele and the Y-axis represents the signal value for the second allele. Guided by the scatter plot appropriate values for five additional parameters were set. The first parameter is called noise allowance, represented in the plot by a green diagonal line. Any sample that is below this line is labelled as low quality and no genotype is assigned.

The remaining four variables correspond to two horizontal and two vertical lines in the plot. The four lines together divide the plot in a three-by-three grid. All data points represented in the upper-left box of the grid are labelled homozygous for allele 2. All data points represented in the lower-right box of the grid are labelled homozygous for allele 1. All data points represented in the centre box of the grid are labelled heterozygous, having a copy of both allele 1 and allele 2. Only the data points not within one of these three boxes are highlighted for manual inspection. If necessary, the automatically assigned genotypes were overruled by manually selecting the preferred allele in the columns neighbouring the automatically assigned genotypes. Selecting "X" will exclude the sample from further analysis.

Finally, SNPtyper summarised all genotypes in a separate worksheet named *Output*. From this output worksheet the data were rearranged and reformatted to meet the layout required for further data processing. Two formats are readily available, including the format for a Crimap input file (<http://www.genome.iastate.edu/~max/lab/CRIMAPwkshp/crimap-doc.html>).

The genotypes assigned using the modified multiplex protocol were validated by resequencing the SNP regions from twelve individuals. For all 12 individuals, all 10 assigned genotypes were in full agreement. The accuracy of the newly developed SNPtyper was assayed by genotyping all 42 individuals of seven families (each consisting of two parents and 4 offspring) for a set of 10 SNP markers. The raw data were analysed both by manual inspection and by SNPtyper. All genotypes assigned by SNPtyper were identical to those assigned manually. Furthermore, the assayed SNP markers showed Mendelian inheritance within these families.

In conclusion, application of a modified multiplex SNaPshot protocol enables generation of over 1000 SNP genotypes per day for a single experimenter or an automated gel-based fluorescent sequencer like the ABI377. Because of the lagging development of analysis software for the Macintosh platform, however, analysis of the generated data was tedious and labour-intensive. The tool presented, SNPtyper, facilitates rapid, accurate



and convenient analysis of this data, while until now the analysis on the Macintosh platform was tedious and labour-intensive.

Both SNPtyper and a Genotyper file containing all settings applied in this set-up are freely available through <http://www.zod.wau.nl/abg/snptyper>.

## References

1. Dietrich, W.F., Weber, J.L., Nickerson, D.A., Kwok, P.Y. (1999) Identification and analysis of DNA polymorphisms, in *Genome analysis, a laboratory manual*, vol. 4: Mapping Genomes (Green, E.D., Birren, B., Klapholz, S., Myers, R.M., Hieter, P. (ed.), Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY, USA. pp.135-186
2. Saiki RK, Walsh PS, Levenson CH, Erlich HA. (1989) Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc Natl Acad Sci U S A* 1989 Aug;86(16):6230-4
3. Hacia JG, Sun B, Hunt N, Edgemon K, Mosbrook D, Robbins C, Fodor SP, Tagle DA, Collins FS. (1998) Strategies for mutational analysis of the large multiexon ATM gene using high-density oligonucleotide arrays. *Genome Res* 1998 Dec;8(12):1245-58
4. Prince JA, Feuk L, Howell WM, Jobs M, Emahazion T, Blennow K, Brookes AJ. (2001) Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation. *Genome Res* 2001 Jan;11(1):152-62
5. Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K. (1995) Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl* 1995 Jun;4(6):357-62
6. Livak KJ. (1999) Allelic discrimination using fluorogenic probes and the 5' nuclease assay. *Genet Anal* 1999 Feb;14(5-6):143-9
7. Tyagi S, Kramer FR. (1996) Molecular beacons: probes that fluoresce upon hybridization. *Nat Biotechnol* 1996 Mar;14(3):303-8
8. Pastinen T, Kurg A, Metspalu A, Peltonen L, Syvanen AC. (1997) Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res* 1997 Jun;7(6):606-14
9. Nikiforov TT, Rendle RB, Goelet P, Rogers YH, Kotewicz ML, Anderson S, Trainor GL, Knapp MR (1994) Genetic Bit Analysis: a solid phase method for typing single nucleotide polymorphisms. *Nucleic Acids Res* 1994 Oct 11;22(20):4167-75

10. Braun A, Little DP, Koster H. (1997) Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. *Clin Chem* 1997 Jul;43(7):1151-8
11. Kristensen, V.N., Kelefiotis, D., Kristensen, T. and Børresen-Dal, A.L. (2001) High-throughput methods for detection of genetic variation. *BioTechniques* 30, 318-332
12. Syvänen, A.-C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* 2, 930-942
13. Syvänen, A.-C. (1999) From gels to chips: "Minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. *Human Mutation* 13, 1-10
14. Lindblad-Toh, K., Winchester, E., Daley, M.J., Wang, D.G., Hirschborn, J.N., Laviolette, J.Ph., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., Shah, N., Thomas, D., Fan, J.B., Gingeras, T., Warrington, J., Patil, N., Hudson, T.J. and Lander, E.S. (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics* 24, 381-386
15. Henegariu O, Heerema NA, Dlouhy SR, Vance GH, Vogt PH. (1997) Multiplex PCR: critical parameters and step-by-step protocol. *Biotechniques* 1997 Sep;23(3):504-11
16. Makridakis NM, Reichardt JK (2001) Multiplex automated primer extension analysis: simultaneous genotyping of several polymorphisms. *Biotechniques* 2001 Dec;31(6):1374-80

## Chapter 5

# **The IGF2-intron3-G3072A substitution explains a major imprinted QTL effect on backfat thickness in a Meishan X European white pig intercross**

Bart Jungerius<sup>1</sup>; Anne-Sophie Van Laere<sup>2</sup>; Marinus te Pas<sup>3</sup>;  
Bernard van Oost<sup>4</sup>; Leif Andersson<sup>2, 5</sup>; Martien Groenen<sup>1</sup>

<sup>1</sup> Department of Animal Breeding and Genetics,  
Wageningen University, Wageningen, Netherlands.

<sup>2</sup> Department of Animal Breeding and Genetics,  
Swedish University of Agricultural Sciences, Sweden.

<sup>3</sup> Animal Sciences Group,  
Wageningen University and Research centre, Lelystad, Netherlands.

<sup>4</sup> Department of Laboratory Animal Sciences,  
Utrecht University, Utrecht, Netherlands.

<sup>5</sup> Department of Medical Biochemistry and Microbiology,  
Uppsala University, Uppsala, Sweden.

Accepted for publication in Genetical Research

### Summary

Several experimental crosses in pigs have been used to detect QTLs for a variety of production traits like fatness and meat quality. A paternally expressed QTL for backfat thickness (BFT) was identified near the *IGF2* locus on the distal tip of SSC2p in three experimental F<sub>2</sub>-populations. Recently, a mutation in a regulatory element of the *IGF2* gene was identified as the Quantitative Trait Nucleotide (QTN) underlying the major QTL effect on muscle growth and BFT in crosses between Large White and Wild Boar or Pietrain. This study demonstrates that the *IGF2* mutation also controls the paternally expressed QTL for backfat thickness in a cross between Meishan X European Whites. In addition, a comparison of QTL for backfat thickness measured by Hennessy grading probe (HGP) and by ultrasound measurement (USM) was made. In the USM analyses, the *IGF2* mutation explains the entire QTL effect on SSC2p, whereas in the BFT-HGP analysis the presence of a second minor QTL can not be excluded. Finally, this study shows that the *IGF2* mutation has no pleiotropic effect on a paternally expressed QTL for teat number identified on SSC2 in the same population.

## Introduction

Several experimental crosses in pigs have been used to detect QTLs for a variety of production traits like fatness and meat quality. For backfat thickness (BFT) significant QTL effects have been detected on 10 different chromosomes (reviewed by Bidanel and Rothschild, 2002). On the short arm of chromosome 2 (SSC2p) a paternally expressed QTL for BFT was identified in several experimental F<sub>2</sub>-populations. The QTL was first mapped near the *IGF2* locus on the distal tip of SSC2p using Large White X Wild Boar (Jeon *et al.*, 1999) and Large White X Pietrain (Nezer *et al.*, 1999) intercrosses. In a Meishan X European White intercross the maximum QTL peak was more proximal at 32 cM, even though the marker within the *IGF2* gene (*SWC9* at 2 cM) was highly informative (de Koning *et al.*, 2000). Therefore, it could not be excluded that the QTL identified in the latter cross was different from the other two reported QTL. It was hypothesised that the observed difference in the QTL position was possibly caused by the presence of two neighbouring QTL: one paternally expressed and one Mendelian additive QTL (Rattink *et al.*, 2000). However, *IGF2* was still within the confidence interval of the QTL in the Meishan X European White cross. Because *IGF2* is known to be paternally expressed, it is the prime candidate gene underlying the paternally expressed QTL for BFT in the Meishan X White cross. Recently, a mutation in a regulatory element of the *IGF2* gene was identified as the Quantitative Trait Nucleotide (QTN) underlying the major QTL effect on muscle growth and BFT in the crosses between Large White and both Wild Boar and Pietrain (Van Laere *et al.*, 2003). The mutation consists of a single nucleotide substitution (G-A) at position 3072 in the third intron of *IGF2* and strong evidence for a causal relation with the observed QTL effect was provided. To analyse the contribution of this mutation to the observed QTLs for BFT in the Meishan X European White cross, this population was genotyped for this mutation. QTL results obtained with two methods of BFT measurement were compared; BFT was estimated using a Hennessy Grading Probe (BFT-HGP), which is measured on the carcass, and by UltraSound measurements (BFT-USM) on live animals. In addition, the QTL analysis for chromosome 2 was repeated to test for the possible presence of a second QTL in the same region.

In the same Meishan X European White population, a paternally expressed QTL for teat number (TN) was identified near *SWC9* on the distal tip of the short arm of SSC2 (Hirooka *et al.*, 2001). Because of the colocalisation and the same mode of inheritance the QTL data for TN were reanalysed to test whether the *IGF2*-intron3 QTN may have pleiotropic effects on teat numbers.

## Materials and Methods

### *Experimental population and phenotypes*

A detailed description of the population used in this study was given previously (Janss et al., 1997a). Briefly, a cross between Meishan and Dutch Large White and Landrace lines was established for the detection of major genes (Janss et al., 1997b) and for QTL analyses (de Koning et al., 1999, 2000; Rattink et al., 2000).

Body weight and backfat thickness were recorded on 1130 F2 pigs from 38 half-sib families shortly before slaughter. The ultrasound measurements (BFT-USM) were performed at 2 to 4 locations at one side of the spine using a Renco Lean-Meater (Renco, Minneapolis, Minnesota, USA). After slaughter, carcass-weight and the backfat thickness, estimated using a Hennessy Grading Probe (BFT-HGP), were recorded on 774 of these F2 pigs. In the Netherlands BFT and muscle content are routinely measured using the Hennessy Grading Probe between the third and fourth rib of the carcass, 6 cm from the spine. For 1173 F2 individuals, the teat number (i.e. the number of morphologically normal teats) was scored.

### *Genotyping*

Thirteen microsatellite markers located on SSC2 were used to obtain genotypes from the F2 animals, their F1 parents and the purebred Meishan grandparents as described by Rattink et al. (2000). Marker names and locations are indicated in figure 1. In addition, all 38 F1 boars and their parents were genotyped for the G-A substitution at position 3072 in the third intron of IGF2. Genotyping was performed by pyrosequencing (Pyrosequencing AB) as described by Van Laere et al. (2003).

### *Extrapolation of the IGF2 SNP to F2 individuals*

The SNP in the *IGF2* gene is located 12.2 kb upstream of *SWC9* in the 2cM interval between *SW2443* and *SWC9*. The SNP genotypes of the F0 and F1 individuals were integrated with the genotypes of a subset of SSC2p microsatellites. (i.e. *SW2443*, *SWC9*, *SW256* and *S0141* located at map position 1, 3, 26 and 40 cM respectively). For each family, parental haplotypes were determined using Simwalk2 (Sobel and Lange, 1996), based on the pedigree information and the inheritance of the microsatellite markers. Simultaneously, the SNP alleles were placed in these haplotypes. The SNP alleles and their parental origin in the F2 individuals were inferred based on the segregation of the haplotype containing the SNP and the flanking markers.

### *QTL analyses*

Statistical analyses were carried out as described in detail by de Koning et al. (1999 and 2000). Briefly, the phenotypic data were adjusted for a number of systematic effects prior to the QTL analysis. The standard set of effects included in the model

consists of sex, weight and a factor to correct for the facility where the measurements were performed (i.e. the five companies that provided the White lines).

Mendelian QTL analyses were based on the line cross concept (Haley et al. 1994), where the original breeds are assumed to be homozygous for different QTL alleles. This model has been extended to test for imprinting (Knott *et al.*, 1998), but a separate test was needed to infer paternal or maternal expression. Therefore the model for imprinting was re-parameterized to test for the direct contribution of the paternally and maternally inherited effect (de Koning *et al.*, 2000). This model facilitates discrimination between QTL showing exclusive paternal expression, exclusive maternal expression or Mendelian expression.

Significance thresholds were determined empirically by permutation with at least 10000 replicates (de Koning *et al.*, 1999). The significance threshold was set at 5% genome-wide risk level (Lander and Kruglyak, 1995). Under all models used the information content is proportional to the variance of the estimators that are used in the regression analyses and has a maximum value of 1.125 (Knott et al., 1998). In this study, however, it is scaled to vary between 0 and 1.

QTL analyses were performed for three different traits: BFT-HGP, BFT-USM and teat number. For each trait, three line-cross QTL analyses were performed each with a different subsets of the population: 1) all 38 half-sib families; 2) the 20 families where the F1 sire was homozygous at the IGF2 QTN (G/G); 3) the 18 families with F1 sires heterozygous at the IGF2 QTN (A/G). In the analyses of BFT-HGP all 774 phenotyped F2 individuals were used, of which 344 with a G/G and 430 with an A/G sire. The same 774 individuals were used in the analyses for HGP-USM analysis. In addition, the BFT-HGP and the BFT-USM data were re-analysed using phenotypic data after correction by the alternative models: the standard model, but with the paternal allele of the *IGF2*-intron3-G3072A mutation as a systematic effect and the standard model, of which the factor weight was omitted.

## Results and Discussion

### *Presence of the IGF2 wild type and mutant alleles*

The *IGF2*-intron3-G3072A mutation was successfully genotyped in 36 of the 38 F1 sires, all 18 of their Meishan sires and 31 of their White dams. All Meishan F0 sires were found to be homozygous for the wild type allele (G/G). In the White F0 dams both the wild type and the mutant allele were represented although they were not equally distributed over the five commercial lines. All F0 dams of the two lines of Dutch Landrace origin were found to be homozygous for the wild type allele (G/G). In the F0 dams from the other three White lines the mutant ('lean' or A) allele was found with frequencies over 80 percent, but it is not fixed in either line. The contrast between absence of the mutant allele and high allele frequencies in all three lines where the mutation is present suggest that once the mutant allele is present within a

population, it reaches high allele frequencies due to positive selection for lean growth. Two out of 38 F1 sires were not genotyped but both the parents were homozygous for the mutation and the genotypes of these F1 sires were deduced. Overall, 20 F1 boars were found to be G/G and 18 were scored as A/G (Table 1).

For the F2 pigs, the genotype for the *IGF2* mutation was not assayed directly, but was inferred based on segregation analysis of flanking microsatellite markers. To do so, the *IGF2* mutation genotypes of the F0 and F1 individuals were combined with the genotypes of the F0, F1 and F2 individuals for four microsatellite markers. Subsequently, these data were used to reconstruct haplotypes that segregate in the population.

Because the *IGF2* mutation is located in the 2 cM interval between microsatellite markers *SW2443* and *SWC9*, a 2% recombination fraction in this marker bracket is expected. Recombinant haplotypes, however, were excluded from further analysis to prevent false allele assignments. Finally, a conclusive paternal allele for the *IGF2* mutation was assigned for 726 F2 pig (94%). For the remainder, the observed segregation of microsatellites could only be explained by the introduction of recombinant haplotypes and were therefore omitted. In total, 175 F2 pigs inherited a mutant (A) allele from their sire and 551 inherited the wild-type (G) allele from their sire.

**Table 1:** Distribution of genotypes at *IGF2*-intron3-nucleotide 3072 among F1 sires in Chinese Meishan X White intercross. The sires are grouped by the white line origin of their mother. No A/A sires were present because all Meishan fathers were G/G. \* Although they may not be purebred, the origin reflects the genetic backgrounds of the F0 sows.

Line	Origin*	G/G	A/G
1	Great Yorkshire	3	5
2	Large White	1	6
3	Large White	1	7
4	Dutch Landrace	7	0
5	Dutch Landrace	8	0
Total		20	18

### *Phenotypic data*

For 774 F2 pigs in 38 half-sib families both live weight and carcass weight were recorded. In this dataset the correlation between live weight and carcass weight was 0.81. For the same 774 F2 pigs both BFT-HGP and BFT-USM were measured. The correlation between BFT-HGP and by BFT-USM was 0.74. The mean BFT-HGP was 22.0 mm. All F2 individuals were grouped based on the SNP-allele inherited from the father. The group of individuals which inherited an A allele from the paternal side had



an average BFT-HGP of 20.8 mm whereas the individuals that inherited a paternal allele G had an average BFT-HGP of 22.4 mm (Table 2).

**Table 2:** Mean backfat thickness of the F2 individuals for both Hennessy Grading Probe (BFT-HGP) and Ultrasound measurements (BFT-USM) (raw data). The groups are based on the *IGF2* QTN allele inherited from the F1 sire.

Group	Number	BFT-HGP (SE)	BFT-USM (SE)
A	344	20.8 (0.33)	14.7 (0.22)
G	430	22.4 (0.26)	15.5 (0.17)
All	744	22.0 (0.20)	15.4 (0.13)

**Table 3:** Summary of QTL results for backfat thickness SSC2 under the paternal expression model.

Measurement <sup>a</sup>	Population	n	Correction <sup>b</sup>	Best position	F <sup>d</sup> ratio	QTL effect <sup>e</sup>	Marker bracket(s) <sup>f</sup>
BFT-HGP	All	774	std	32	27.9 <sup>***</sup>	0.04	SW2443-S0091
	AG	344	std	6	34.6 <sup>***</sup>	0.09	SW2443-SW240
	GG	430	std	64	6.8	0.02	
BFT-USM	All	774	std	4	19.6 <sup>**</sup>	0.02	SW2443-S0141
	AG	344	std	3	25.6 <sup>**</sup>	0.07	SW2443-S0141
	GG	430	std	150	6.3	0.01	
BFT-HGP	All	774	std+IGF2	40	13.1 <sup>*</sup>	0.02	
BFT-USM	All	774	std+IGF2	8	7.6	0.01	

<sup>a</sup> BFT-HGP: backfat thickness by Hennessy Grading Probe, BFT-USM: backfat thickness by ultrasound measurement

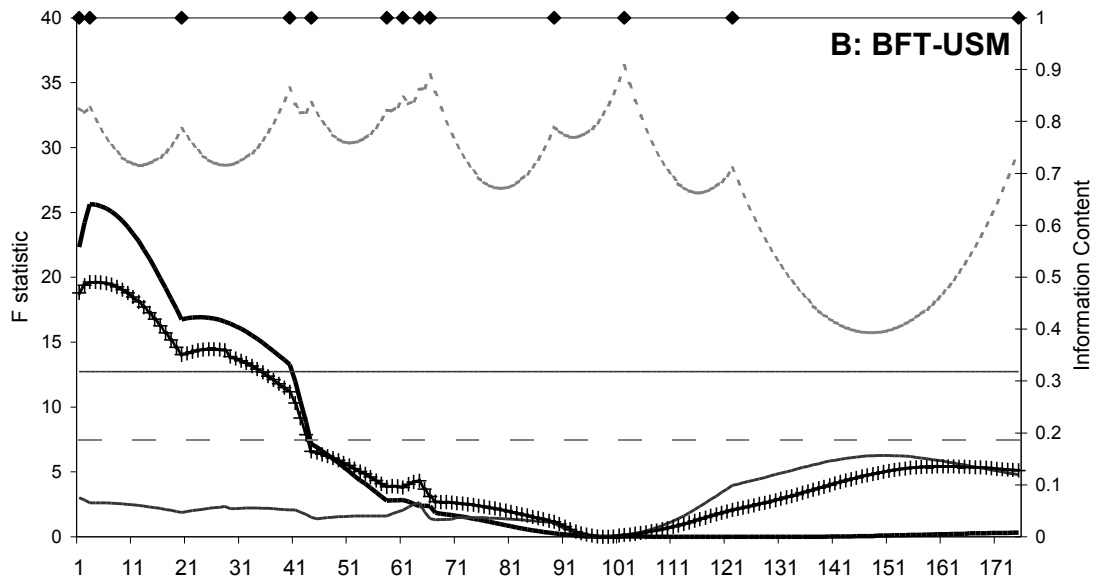
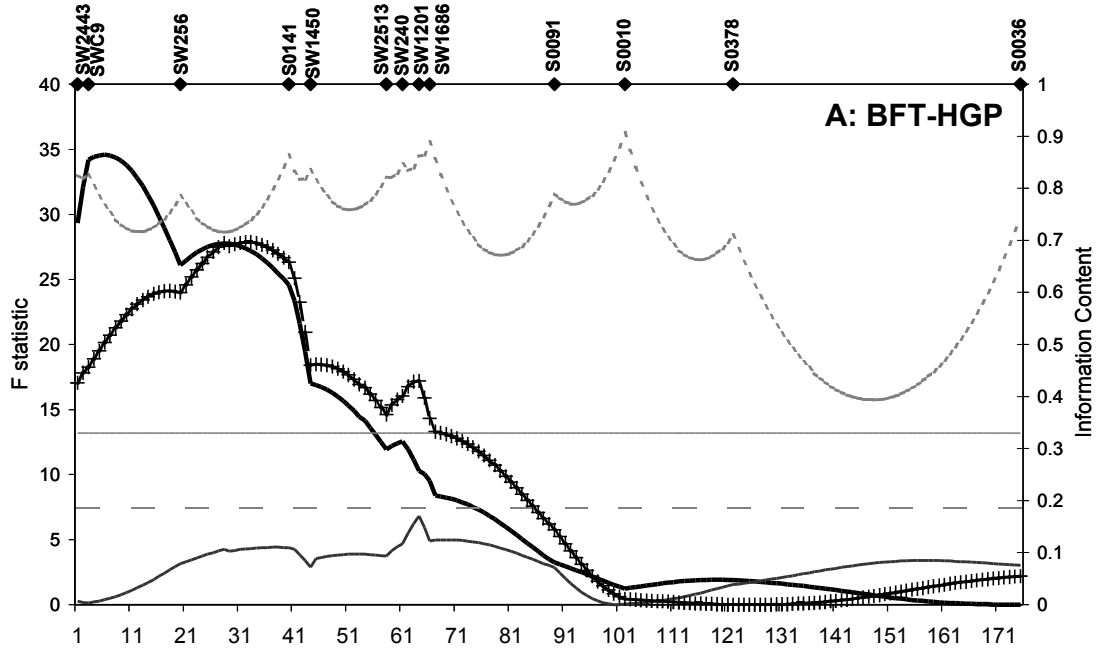
<sup>b</sup> std (weight + sex + facility), +IGF2: including the paternal allele of *IGF2*.

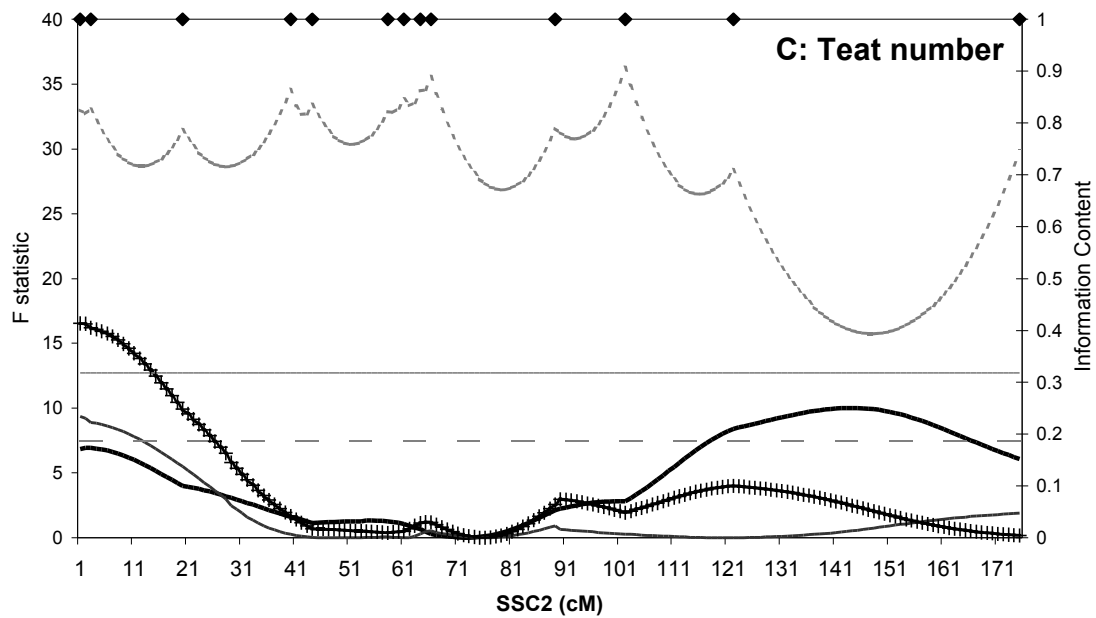
<sup>c</sup> Mend: Mendelian, Pat: Paternally expressed) ;

<sup>d</sup> P: † p<0.10 , \* p<0.05, \*\* p<0.01, \*\*\* p<0.001 (based on tabulated F values after 10000 permutations).

<sup>e</sup> fraction of observed variation explained by the QTL.

<sup>f</sup> Marker bracket(s) were F statistic exceeds genome-wide significance level.





**Figure 1:** QTL analyses on SSC2 for BFT by Hennessy Grading Probe (A), for BFT measured by UltraSound measurement. (B) and for teat number (C). The thick solid line represents the QTL from the segregating (A/G) families. The thin solid line is from non-segregating (G/G) families. The '+'-marked line is a combined analysis of segregating and non-segregating families. Information content is represented by a grey dotted line. Horizontal lines indicate the significance (solid) and suggestive (dotted) levels. Marker names and positions are indicated at the top.

### *BFT QTL analyses and the IGF2 mutation*

All 774 phenotyped individuals were used in the analyses of BFT-HGP. The same 774 individuals were used in the analyses for BFT-USM to allow comparison of the results, although at the cost of reduced population size and lower F statistics. Data were also analysed using phenotypic data after correction by an alternative model in which the paternal IGF2 allele is included. The results of the QTL analyses with the paternal expression model are summarised in Table 3. QTL analyses following the Mendeian model were also performed, but in all occasions the F statistics were lower than with the paternal expression model (data not shown).

Initially, one QTL analysis for BFT-HGP was performed which included all 38 families. As expected, the outcome closely resembled the QTL reported by de Koning et al. (2000). Since *IGF2* is paternally expressed the QTL effect is expected to segregate only in families founded by F1 sires that are heterozygous A/G at the QTN. In a joint QTL analysis of the 20 families of which the F1 sire was found to be homozygous for the IGF2 wild type allele (G/G) the F statistic did not reach the genome-wise suggestive threshold along the entire chromosome. In a similar analyses of the 18 families of which the F1 sire was found to be heterozygous for the mutation (A/G) the F statistic exceeded the genome-wise significant level for both BFT-HGP and BFT-USM (Figure 1A and B). These results clearly demonstrate that the IGF2-intron3-G3072A substitution splits the population in two groups; families segregating for a major QTL and families that do not. From this, it can be concluded that the *IGF2* QTN explains the major part, if not all, of the observed paternally expressed QTL for BFT on SSC2 in this pedigree. This conclusion is supported by the differences in average BFT between individuals which inherited a G (22.4 mm) or an A (20.8 mm) from their father.

### *Additional QTL and influence of measurement techniques*

In addition to the analyses of BFT-HGP data, similar analyses were performed for BFT-USM (Figure 1B). The results of the QTL analyses for BFT-HGP and for BFT-USM are highly consistent for analyses of the 18 families in which both alleles of the *IGF2* SNP are segregating. Although the QTL profiles largely overlap in analyses including all 38 families, the best positions of the QTL peaks for BFT-HGP and for BFT-USM differ almost 30 cM (figure 1A and B, dotted lines).

Based on previous analyses of the BFT-HGP QTL on SSC2 (Rattink et al., 2000; de Koning et al., 2000), it was suggested that the QTL identified in the Meishan X White cross could be different from the QTL reported by Nezer *et al.* (1999) and Jeon *et al.* (1999). A possible explanation for the observed difference in the QTLs best positions was that it might be caused by the presence of two neighbouring QTL: one paternally expressed and one mendelian additive QTL (Rattink et al., 2000). Although no convincing evidence for the presence of a second QTL was found, its presence could not be excluded either.

Now the mutation underlying the primary QTL in the region is known, it is possible to incorporate the paternal IGF2 allele (A or G) as a fixed effect in the phenotypic data. Thus, the presence of a second neighbouring QTL can be validated.

The QTL for BFT-USM showed no evidence for an additional (mendelian) QTL contributing to the phenotype. However, a paternally expressed QTL around 40 cM was indicated in the analysis of the HGP data. Although this analysis is based on phenotypic data that were corrected for a deduced *IGF2* genotype it exceeds the genome-wise significance threshold and the presence of an additional QTL can not be excluded. Because the possible second QTL is only observed with BFT-HGP, this would be a QTL for a characteristic that is measured exclusively by BFT-HGP and not by BFT-USM, e.g. the third layer of backfat. Total backfat consists of three layers of which the third (inner) layer develops at a later stage of growth and often is underdeveloped in European commercial lines. Very thin third layer fat might be difficult to detect by BFT-USM. Consequently, in some cases BFT-USM only represents the thickness of two layers of fat, while BFT-HGP measures the total amount of fat in all three layers.

Another difference between the BFT-HGP and the BFT-USM QTL analyses are corrections for body weight: carcass weight is used for BFT-HGP and live weight for BFT-USM. To eliminate the effect of the difference between carcass weight and live weight, the QTL were re-analysed using phenotypic data that were corrected using a model that did not include body weight. Although for both BFT-HGP and BFT-USM the simplified model (excluding weight) led to a reduced F statistic the F values exceeded the genome-wise significant threshold in all analyses. The patterns of the QTL did not change substantially by omitting the correction for weight, or by correcting BFT-HGP for live-weight and BFT-USM for carcass-weight (data not shown). These findings suggest that the difference in positioning of the peak is not caused by the difference in body weight measurement.

Finally, the shift in QTL position between BFT-HGP and BFT-USM might be explained by a ghost QTL at position 40 cM. This ghost QTL might be the result of the differences in F-values at the *IGF2* locus between BFT-HGP and BFT-USM for the non-segregating families. In these families the F-value is absolutely zero for BFT-HGP but around 5 for BFT-USM (figures 1A and B, thin solid lines).

Despite the observed differences, both BFT-HGP and BFT-USM appear to be equally efficient in picking up the major QTL effect in segregating families.

#### *Teat number QTL analyses and the IGF2 mutation*

A total of 1173 F2 individuals in 38 half-sib families have previously been applied to perform a QTL scan for teat number resulting in the identification of a paternally expressed QTL at the distal tip of SSC2p (Hirooka et al., 2001). A QTL with an effect on teat number was observed both in the analyses of the 20 families of which the F1 sire was found to be homozygous for the *IGF2* wild type allele (G/G) and in the analyses of the 18 families of which the F1 sire was found to be heterozygous for the

mutation (A/G) (figure 1C). This shows that the IGF2-intron3-G3072A substitution does not have a pleiotropic effect on teat number.

In conclusion, a paternally expressed QTL for lean meat growth and for BFT was identified on the distal tip of SSC2p in three experimental populations. Recently, the IGF2-intron3-G3072A substitution was reported to be the underlying causative mutation for this QTL (Van Laere et al., 2003). This study demonstrates that the *IGF2* mutation also causes the paternally expressed QTL for backfat thickness in the cross between Meishan X European Whites. Although differences between QTL for BFT-HGP and BFT-USM exist, both methods are caused by the same mutation and both methods are equally efficient to detect the observed QTL, although in the BFT-HGP analysis the presence of a second minor QTL (at 40 cM) can not be excluded. The QTL for teat number identified on SSC2 in the Meishan X European White cross (Hirooka et al., 2001) was not affected by the previously described *IGF2* mutation. Thus, another mutation in the IGF2 region may cause this effect on teat numbers.

### Acknowledgements

We thank Henk Bovenhuis, Richard Crooijmans, Bart Ducro, Piet de Groot, Dirk-Jan de Koning, Jan van der Poel, Annemieke Rattink and Maria Siwek for their assistance, help and suggestions.

### References

- Bidanel, J.P., and M. Rothschild (2002). Current status of quantitative trait locus mapping in pigs. *Pig News and Information*. 23(2):39N-54N.
- de Koning DJ, Janss LL, Rattink AP, van Oers PA, de Vries BJ, Groenen MA, van der Poel JJ, de Groot PN, Brascamp EW, van Arendonk JA (1999). Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (*Sus scrofa*). *Genetics* **152**(4), 1679-90.
- de Koning DJ, Rattink AP, Harlizius B, van Arendonk JA, Brascamp EW, Groenen MA. (2000). Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proceedings of the National Academy of Sciences of the United States of America*. **97**(14), 7947-50.
- Haley CS, Knott SA, Elsen JM (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*. **136**(3), 1195-207.
- Hirooka H, de Koning DJ, Harlizius B, van Arendonk JAM, Rattink AP, Groenen MAM, Brascamp EW, Bovenhuis H. (2001) A whole-genome scan for quantitative trait loci affecting teat number in pigs. *Journal of Animal Science* **79**, 2320-26.

Janss LL, van Arendonk JA, Brascamp EW. (1997a) Bayesian statistical analyses for presence of single genes affecting meat quality traits in a crossed pig population. *Genetics*. **145**(2), 395-408.

Janss LL, Van Arendonk JA, Brascamp EW. (1997b) Segregation analyses for presence of major genes affecting growth, backfat, and litter size in Dutch Meishan crossbreeds. *Journal of Animal Science*. **75**(11), 2864-76.

Jeon JT, Carlborg O, Tornsten A, Giuffra E, Amarger V, Chardon P, Andersson-Eklund L, Andersson K, Hansson I, Lundstrom K, Andersson L. (1999) A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nature Genetics* **21**(2), 157-8.

Knott SA, Marklund L, Haley CS, Andersson K, Davies W, Ellegren H, Fredholm M, Hansson I, Hoyheim B, Lundstrom K, Moller M, Andersson L. (1998) Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics*. **149**(2), 1069-80.

Lander E, Kruglyak L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*. **11**(3), 241-7.

Nezer C, Moreau L, Brouwers B, Coppeters W, Dettileux J, Hanset R, Karim L, Kvasz A, Leroy P, Georges M. (1999). An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nature Genetics*. **21**(2), 155-6.

Rattink AP, De Koning DJ, Faivre M, Harlizius B, van Arendonk JA, Groenen MA. (2000) Fine mapping and imprinting analysis for fatness trait QTLs in pigs. *Mammalian Genome*. **11**(8), 656-61.

Sobel E, Lange K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*. **58**(6), 1323-37.

Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, Andersson G, Georges M, Andersson L (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature*. **425**(6960), 832-6.





## **Chapter 6**

### **Estimation of the extent of linkage disequilibrium in seven regions of the porcine genome**

Bart Jungerius<sup>1</sup>, Jingjing Gu <sup>1</sup>, Richard Crooijmans<sup>1</sup>, Jan van der Poel<sup>1</sup>, Bernard van Oost <sup>2</sup>, Marinus te Pas <sup>3</sup>, Martien Groenen<sup>1</sup>

<sup>1</sup> Department of Animal Breeding and Genetics, Wageningen University, Wageningen, Netherlands.

<sup>2</sup> Department of Laboratory Animal Sciences, Utrecht University, Utrecht, Netherlands.

<sup>3</sup> Animal Sciences Group, Wageningen University and Research centre, Lelystad, Netherlands.

Submitted

### **Abstract**

Linkage disequilibrium (LD) refers to the correlation among neighboring alleles, reflecting non-random patterns of association between alleles at (nearby) loci. A better understanding of LD in the porcine genome is of direct relevance for identification of genes and mutations with a certain effect on the traits of interest.

Here, 215 SNPs in 7 genomic regions were genotyped in individuals of three breeds. Pairwise linkage disequilibrium was calculated for all marker pairs. To estimate the extent of LD, all pairwise LD values were plotted against the distance between the markers. Based on SNP markers in four genomic regions analyzed in three panels from populations of Large White, Dutch Landrace and Meishan origin, useful LD is estimated to extend for approximately 40 to 60 kb in the porcine genome.

## Introduction

Many economically important traits in livestock species are quantitative traits. Some of the genes that contribute to these traits have been localized using quantitative trait locus (QTL) mapping. Because the co-segregating piece of DNA can only be delimited by observation of recombination events between a marker and the causative locus itself, the QTL can only be assigned to a large interval. Typically, the number of informative meioses in pedigrees for linkage mapping typically is limiting. An additional method to narrow down the mapping interval is based on the analysis of linkage disequilibrium (LD). LD mapping relies on segregation of variations in natural populations, which represent more meioses and historical recombination events and therefore can yield higher resolution maps. Consequently, LD mapping can be applied to narrow down a QTL region or it can be applied in whole-genome-association studies to search for genes involved in quantitative traits. An essential issue in the LD mapping, however, is the marker density required for this strategy. This density strongly depends on the extent of LD, which refers to the distance at which markers still are in sufficient LD for mapping purposes.

In human data, both simulations and empirical data have shown that LD may extend for somewhere between 5 and 500 kb (Clark et al., 1998; Rieder et al., 1999; Moffatt et al., 2000; Templeton et al., 2000; Dunning et al., 2000). The extent of LD is not equally distributed over the genome and LD patterns vary between populations (Goddard et al., 2000; Kidd et al., 2000; Reich et al., 2000).

Despite the recent developments in LD research in human, there is only limited information available on LD or the extent of LD in the porcine genome. A better understanding of LD in the porcine genome is of direct relevance for identification of genes and mutations with a certain effect on the traits of interest.

In this study, 200 SNP markers from seven regions of 80 to 160 kb long were genotyped. Four of these regions originate from a 2.5 Mb genomic region on SSC18q24. The other regions originated from SSC18q, SSC3 and for one its map position is not known. In total, genotypes were derived for 112 individuals of three breeds, including Meishan, and two commercial Dutch lines of predominant Large White and Dutch Landrace background.were genotyped

## Materials and Methods

### *Population*

In total, 63 individuals in three panels were used to calculate LD. The first panel (panel M) consisted of 18 sires of a Meishan population that was kept in the Netherlands. The population is expected to be inbred because of a limited population size that was kept constant over several years and generations. The second panel consisted of 27 dams of a commercial Dutch line of predominantly Large White origin (panel LW). The third panel consisted of 18 dams of a second commercial Dutch line of predominantly Dutch Landrace origin (panel LR). Within each panel, the subjects were selected to be as unrelated as possible. All three panels were part of the founder generation of an experimental intercross population between Meishan and Dutch White lines that has been described in detail by Janss et al. (1997). To enable pedigree-based haplotype reconstruction, 49 samples of intercross offspring were genotyped. In total, the 112 selected samples comprised 18 separate pedigrees, each consisting of one Meishan sire, 1 to 4 White dams and 1 to 5 intercross offspring.

### *Genomic regions*

Seven porcine BAC clones with finished sequences were selected from Genbank. All seven BAC clones originated from porcine genomic regions that show synteny with human chromosome 7 (HSA7). Four of these BAC clones are part of a 2.5 Mb BAC contig originating from porcine chromosome 18 (SSC18). The BAC contig and its BAC clones were retrieved from the website of the NIH Intramural Sequencing Center ([www.nisc.nih.gov](http://www.nisc.nih.gov)). Two BACs originate from SSC 3, 18 and for one BAC the origin is unknown. For convenience, we labeled the genomic regions A through F (Table 1).

To identify SNPs in the regions at variable pairwise distances, PCR primers were designed for amplification of subregions of around 700 bp with target positions at 0, 2, 5, 10, 20, 40, 80 and 160 kb in the BAC sequences. For BACs E, F and G, however, only PCR products at both ends of the insert were used.

### *PCR and sequencing*

PCR reactions were carried out in 24 µl volumes. Reactions contained 180 ng DNA, 1.5 mM MgCl<sub>2</sub>, 50 mM KCl, 10 mM Tris-HCl pH 8.3, 1 mM tetramethylammoniumchloride (TMAC), 0.1% Triton X-100, 0.01% gelatine, 200 µM of each dNTP, 0.5 unit Silverstar DNA polymerase (Eurogentec, Seraign, Belgium), 6 pmoles of each primer, and were covered with 12 µl mineral oil. Reactions were performed in a PTC100 thermocycler (MJ Research, Watertown, MA, USA) with a 5 min initial denaturation at 95°C followed by 35 cycles of 30 s at 95°C, 45 s at annealing

temperature and 60 s at 72°C, and a final extension for 10 min at 72°C. The annealing temperatures were optimized for each primer pair and varied between 50°C and 65°C. After PCR the excess of primers was removed by running the samples over MANU030 PCR columns in 96-well format (Millipore, USA). From each sample, 2 µl was checked on agarose gel to estimate the DNA concentration. Sequencing reactions were performed with either the forward or the reverse amplification primer. Cycle sequencing reactions contained 100-400 ng of purified PCR product, 2 µl of Big Dye Terminator Rtmix (PE, Foster City, CA, USA), 2 µl of Half Big Dye Buffer (Genetix, New Milton, UK) and 0.8 pmol of either primer in a final volume of 10 µl. Excess dye terminator was removed by purification of the samples on Montage SEQ96 columns in 96-well format (Millipore, USA). After purification, the DNA was dissolved in 10 µl Injection Solution (Millipore, Bedford, MA, USA) and mixed with 10 µl formamide and denatured for 3 min. at 95°C before analysis on an ABI 3100 automated sequencer (ABI, Foster City, CA, USA).

**Table 1:** BAC clones used, their accession numbers, lengths and map positions in pig and human.

Region	Clone name	Accession number	Insert length (kb)	Porcine position	Human position
A	RP44-348C20	AC091506	195	3p15-p17	7q11-7q22
B	RP44-418J07	AC087160	167	18q	7q31
C	RP44-349L07	AC091723	77	18q24	7p14-p15 <sup>1)</sup>
D	RP44-428N16	AC129961	154	unknown	7
E	RP44-74O11	AC091404	102	18q24	7p14-p15 <sup>1)</sup>
F	RP44-397B1	AC091507	174	18q24	7p14-p15 <sup>1)</sup>
G	RP44-524D1	AC096852	69	18q24	7p14-p15 <sup>1)</sup>

<sup>1)</sup> BACs E, F, C and G are part of a BAC contig generated by the NIH Intramural Sequencing Center ([www.nisc.nih.gov](http://www.nisc.nih.gov)).

#### *Polymorphism identification*

The SNP discovery panel consisted of 16 individuals: 6 Meishan boars and 5 sows of each of the White lines. Trace files were analyzed using the Pregap4 program of the Staden software package (Bonfield et al., 1996; <http://www.mrc-lmb.cam.ac.uk/pubseq>). Using Pregap4, all chromatograms were analyzed: bases were called with Phred (Ewing et al., 1998a, 1998b) and low quality regions were masked. Next, all sequences passing Pregap4 were entered in a Gap4 database. In the Gap4 program (Bonfield et al., 1995)

a normal shotgun assembly was performed with a minimal initial match of 20 bp, a maximum of 25 pads per read and a maximum of 5 % mismatches. If necessary, the assembly was finished manually using the Gap4 *Join contig* interface. Positions at which a disagreement occurred were highlighted in the Gap4 contig editor window using the *Highlight disagreements* option. All nucleotide positions at which a disagreement occurred were tagged as a putative SNP position.

### *SNP Genotyping*

After identification of SNPs from the 16 individuals, the remaining 96 samples were genotyped for these SNPs by either single-base extension (SBE) or resequencing. The choice for SBE or resequencing was based on the number of SNPs in the amplified region. SBE reactions were typically carried out in 16-plex format. In case several SNPs were to be genotyped from one PCR product, however, it is more efficient to resequence all 96 individuals.

SBE genotyping was performed using the multiplex SNaPshot kit (ABI, Foster City, CA, USA) following the protocol provided by the manufacturer. Basically, SBE primers were designed to have a specific 3'-end 18-25 bp in length. A non-specific 5'-tail was used to create primers 25-120 bp in length, with 5 bp intervals. SNaPshot products were analyzed on an ABI3100 automated sequencer. Resulting data were analyzed using the GeneMapper software package (v3.0 ; ABI) and checked by two independent persons. In addition, all marker data were checked for Mendelian inheritance within each sire-dam-offspring trio. If for a marker, the alleles within a trio did not follow Mendelian inheritance, the marker was excluded from further analysis for this trio. Resequencing was performed as described above, although usually the products were sequenced in either forward or reverse direction. All reads were grouped per PCR product and analyzed using POSA, a sequence analysis pipeline that provides base calling, contig assembly and annotation of polymorphic positions using Phred, Phrap and Polyphred (Aerts et al., submitted).

### *Haplotyping*

For all 18 pedigrees haplotype sets were reconstructed for each of the seven target regions using SIMWALK2 (Sobel et al., 1996). SIMWALK2 ignores marker positions at which less than 75% of the genotypes is assigned. In some cases, SIMWALK2 is not able to reach a satisfactory conclusion and it will prompt for a rerun. After a maximum of three runs, all derived haplotype vectors were retrieved. Pedigrees for which additional runs were still required were omitted and thus ambiguous haplotype vectors are ignored from further analysis. The resulting set of founder haplotype vectors is composed of sets of haplotypes from all three original panels. From this list, the haplotypes were divided in the three groups based on their origin (i.e. M, LW or LR).

To estimate long-range LD, similar analysis were performed on combined data from the BAC contig containing target regions E, F, C and G. Relative positions of the regions in the contig were estimated from the syntenic mouse genome sequence represented by NISC ([www.nisc.nih.gov](http://www.nisc.nih.gov)).

To estimate LD between unlinked loci, similar analysis were performed on combined data of BACs A, D, F and C. The distance between A and D and between D and F were set to match a recombination fraction of 0.5, corresponding to unlinked chromosomal fragments.

### *Pairwise Disequilibrium*

Using the founder haplotype sets identified by SIMWALK2, two measures of LD were calculated:  $D'$  and  $r^2$ . Both measures are based on Lewontin's  $D$  and both are normalized to a scale 0 to 1, allowing comparison of LD values over different marker pairs. Dividing  $D$  by its maximum possible value with regard to the allele frequencies yields  $D'$ . Dividing the squared value of  $D$  ( $D^2$ ) by the product of all four allele frequencies yields  $r^2$ . Marker positions at which the minor allele was found at frequencies below 4% were excluded from LD calculations to prevent artifacts induced by low frequency alleles.

**Table 2:** The numbers of primer pairs, product lengths and the SNP discovery results per target region..

BAC	No. of primer pairs	Total product length (bp)	No. of SNPs	SNP density (bp/SNP)	Diversity index (bp/SNP)
A RP44-348C20	7	4459	35	68	273
B RP44-418J07	4	2374	18	132	531
C RP44-349L07	5	3014	45	67	270
D RP44-428N16	8	4612	53	87	350
E RP44-74O11	1	652	9	72	392 <sup>1)</sup>
F RP44-397B1	2	1272	31	41	222 <sup>1)</sup>
G RP44-524D1	2	1387	24	58	313 <sup>1)</sup>
Total	29	17770	215	83	336

1) SNPs were discovered from a panel of 63 individuals

## Results

### *Sequencing and SNP identification*

In total, 29 PCR products were sequenced from a panel of 16 individuals. All products were 500 to 700 bp long and each product contained at least one SNP. In the total product length of 17.7 kb, 215 SNPs were identified, an overall average of one SNP per 83 bp. Details per region are presented in Table 2. After normalization for the assayed sample size, this corresponds to a nucleotide diversity index (Nei and Li, 1979) of one SNP every 336 bp, consistent with previous studies (Fahrenkrug et al., 2002; Jungerius et al., 2003).

### *Haplotyping*

The 18 pedigrees used in this study comprised 63 founders (126 founder haplotypes) and included a total of 98 meioses. Assuming that 1 Mb accounts for 1 cM, the expected number of recombinant haplotypes in the 75 to 170 kb target regions ranges from 0.07 to 0.17. For region C, 124 founder haplotypes (out of 126) were identified with no recombinants in any of the 18 pedigrees. For each of the other target regions, 17 pedigrees led to identification of 114 through 118 founder haplotypes without recombinants. In all six regions, the founder haplotypes for the remaining pedigree could not be resolved without introduction of recombinant haplotypes. Considering the low number of recombinants expected, these pedigrees were excluded from further analyses to prevent introduction of false non-existing haplotypes.

Haplotypes were reconstructed in a combined analysis of target regions E, F, C and G, which all are part of a 2.5 Mb BAC contig. Consequently, the data set with 98 meioses is expected to contain approximately 2 recombinant haplotypes. After 3 consecutive runs, SIMWALK2 successfully reconstructed non-recombinant haplotypes for 16 out of 18 pedigrees. Again, the two remaining pedigrees were excluded from further analysis.

### *LD versus distance*

The relationship between LD and the physical distance between a pair of markers within the target regions is shown in Figure 1. Together, the seven analyzed genomic regions include 559 pairs of marker separated by less than 160 kb. Because of the exclusion of markers with fixed alleles and markers with low minor allele frequencies (< 4%), the numbers of informative marker pairs per breed were lower: 244 pairs for LW, 280 pairs for DL and 195 pairs for M.

A way to assess the usefulness of LD for mapping purposes is to assume that one of the markers (of the pair) is the causative polymorphism, and to ask what increase in sample size would be required for detection of association with the other marker. Here, a three-



fold increase in sample size was considered to be feasible and consequently  $r^2$  values of .33 or higher represent useful LD. (Kruglyak, 1999; Ardlie et al., 2002).

Typically, the fraction of marker pairs with useful LD decreases with distance, but even at short distances not all marker pairs show useful LD values. For both the White lines (LW and LR) the majority of physically close marker pairs show  $r^2$  values of .33 or greater and are therefore considered to be useful in LD mapping. For LW (figure 1A), 76 (51%) of the pairs at distances of <40kb are in useful LD. Marker pairs with distances between 40 and 80 kb do not show useful LD. Of the marker pairs spanning over 100 kb only 6 (10%) show useful LD. It should be noted, however, that these pairs all share one marker in common and therefore are not completely independent. In LW, complete LD ( $r^2=1$ ) is only observed for distances up to 20 kb.

For LR (figure 1B), the pattern is similar to that of LW: 106 (60%) of the pairs at distances of less than 40 kb are in useful LD. For pairs at distances between 40 and 160 kb only 6 pairs (7%) exceed the threshold of 0.33. In LR, complete LD ( $r^2 = 1$ ) is common over distances up to 40 kb, but is occasionally observed even at approximately 160 kb.

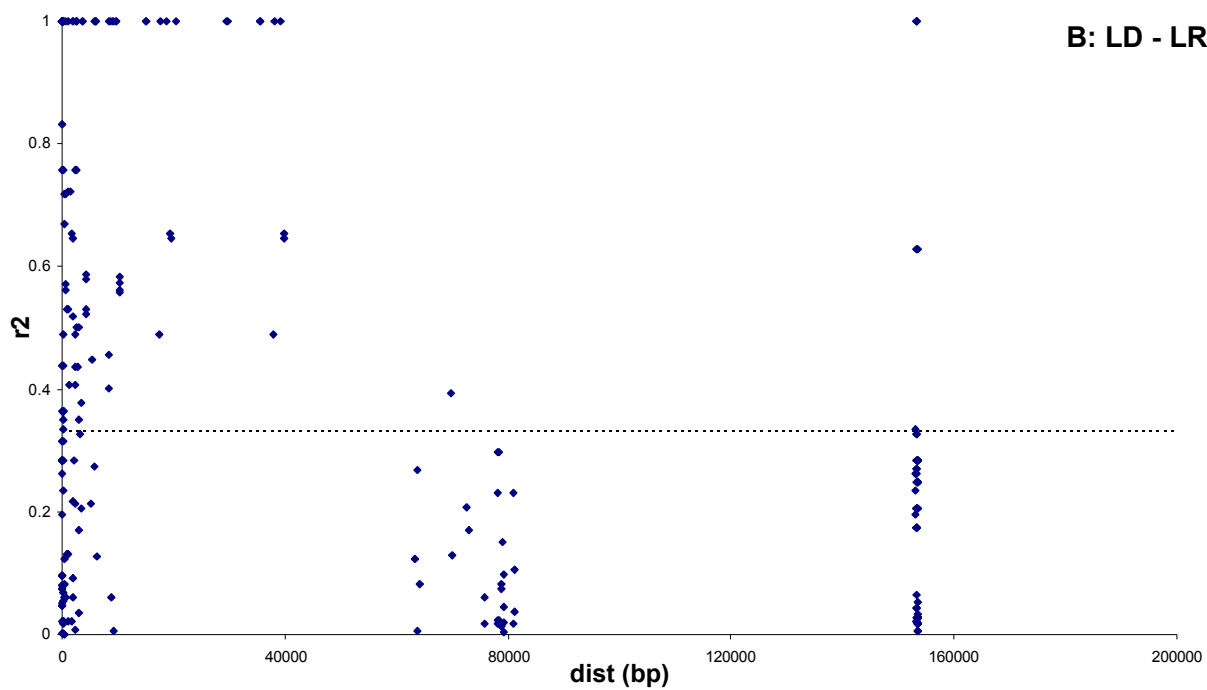
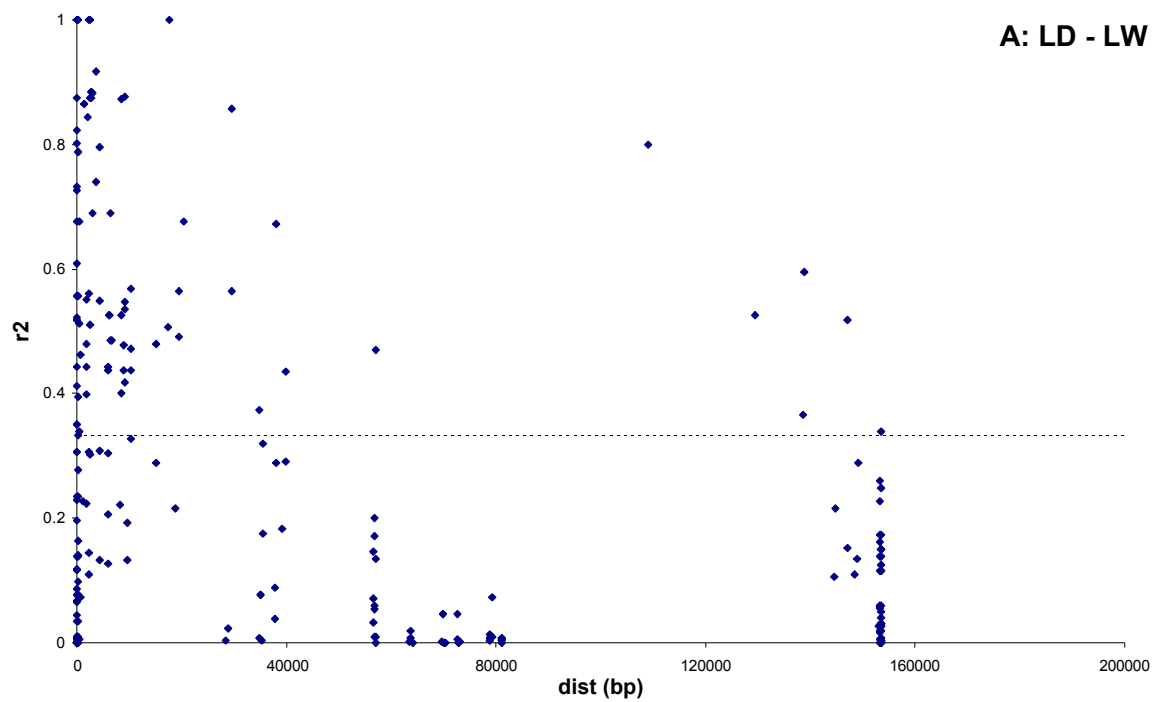
For Meishan, the fraction of marker pairs in useful LD varies between 24% at distances under 40 kb and 12% at distances between 40 and 160 kb. Although the overall fraction of pairs with useful LD is only 19%, it should be noted that pairs in complete LD ( $r^2=1$ ) are detected at any distance, including 160kb.

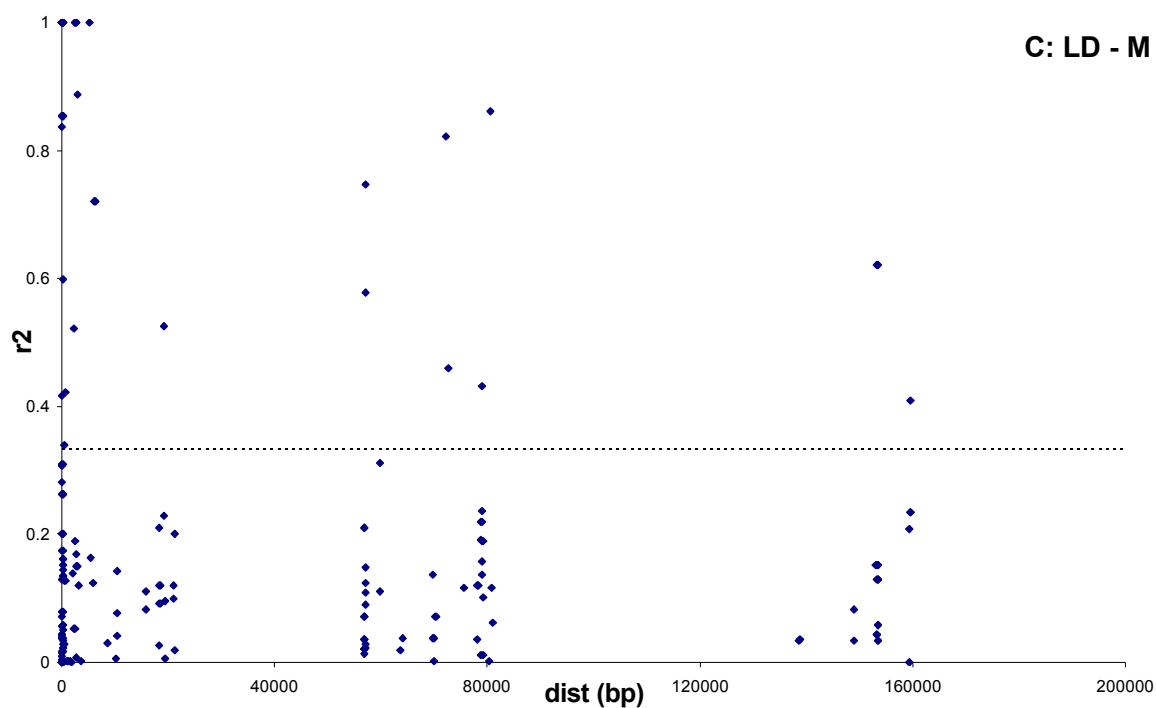
#### *Long-range LD*

To detect the possible presence of long-range LD, the relationship between LD and the physical distance between a pair of markers in the 2.5 Mb BAC contig containing target regions E, F C and G were analyzed. Data of this analysis are shown in Figure 2. For LW, 21 (4.6%) marker pairs spanning 160 to 1700 kb showed useful LD (Figure 2A). For LR, 14 (3.7%) marker pairs spanning 160 to 1700 kb showed useful LD. In the Meishan panel 63 (13.6 %) of the marker pairs spanning 160 to 1700 kb showed useful LD.

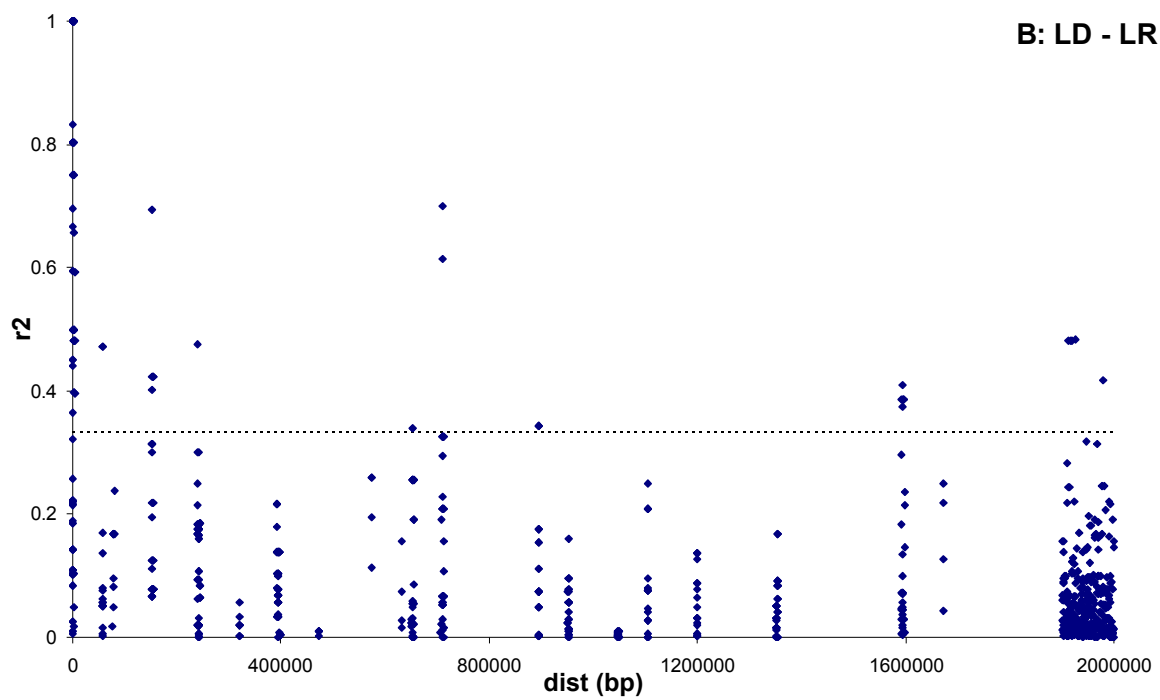
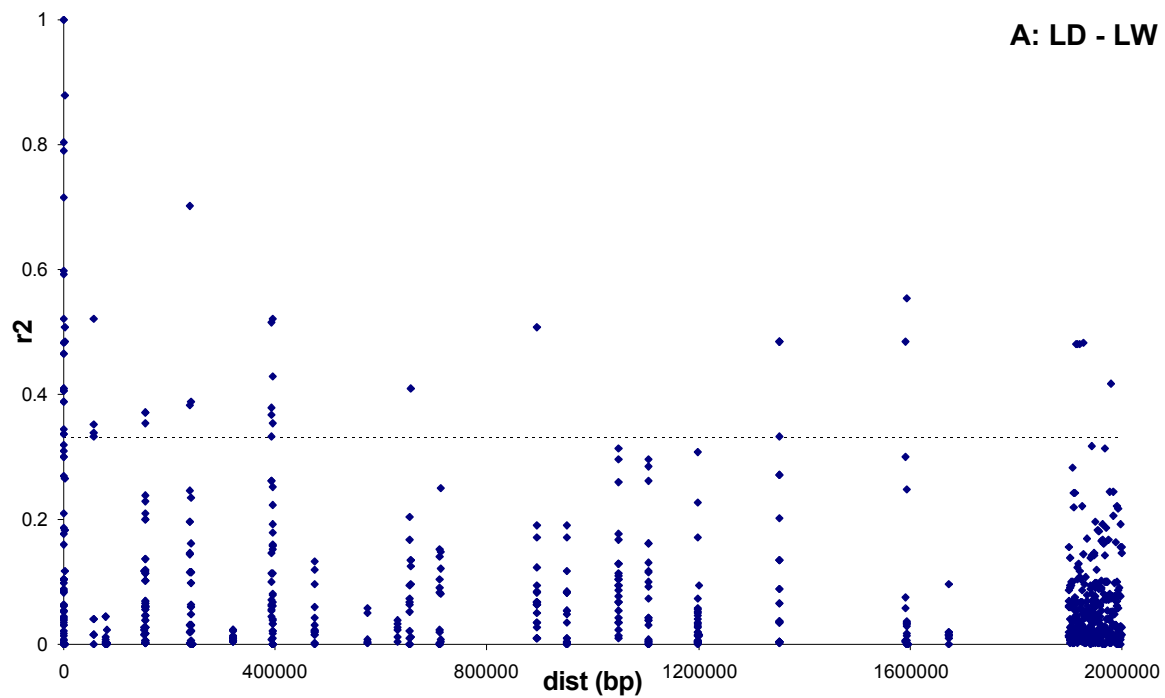
#### *LD between physically unlinked loci*

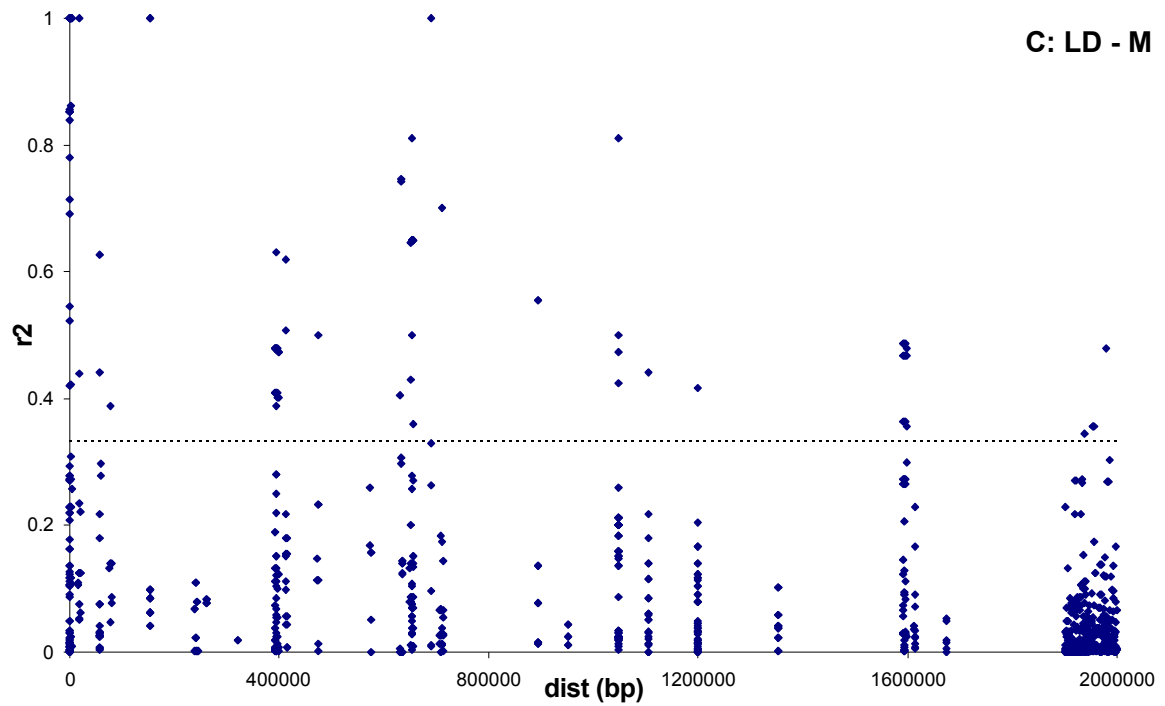
A joint analysis of unlinked loci from target regions A, D and E-F yielded a set of over 400 marker pairs of physically unlinked markers. LD data on the unlinked marker pairs is represented in Figure 2. Most pairs of unlinked markers do not show useful LD in any of the three population panels. For each of the three breeds, however, 5 to 10 (1 to 2.5%) pairs of physically unlinked markers represented useful LD.





**Figure 1:** LD versus physical distance between SNPs for three population panels: Large White (A), Dutch Landrace (B) and Meishan (C). For all three panels, data represented originate from 6 distinct BACs representing different genomic regions.





**Figure 2:** LD versus physical distance between SNPs for three population panels: Large White (A), Dutch Landrace (B) and Meishan (C). For all three panels, data represented originate from four 70-170 kb regions, which are part of a 2500 kb BAC contig derived from SSC18. Data points marked as unlinked represent LD values for pairs of markers that originate from regions that are physically unlinked.

## Discussion

### *Measures of LD.*

In literature several measures of LD have been described, most of which are derived from Lewontin's  $D$  (reviewed by Devlin and Risch, 1995). Here, two commonly used measures,  $D'$  and  $r^2$ , are discussed.  $D'$  is frequently used and it has the same range of values regardless of the frequencies of the SNPs compared. It varies between 0 (complete linkage equilibrium; LE) and 1 (complete LD), but for intermediate values of  $D'$  there is no clear interpretation. Recently, an alternative measure of LD has emerged as the measure of choice for quantifying and comparing LD in the context of mapping:  $r^2$ , sometimes denoted as  $\rho^2$  (Pritchard et al., 2001; Weiss and Clark, 2002).  $r^2$  reflects the correlation of alleles at two loci and is formed by dividing  $D^2$  by the product of the four allele frequencies. Values of  $r^2$  equal to 1 indicate perfect LD and values of  $r^2$  equal to 0 indicate complete linkage equilibrium (as with  $D'$ ). In contrast to  $D'$ , intermediate values of  $r^2$  can be interpreted. Consider a two locus model where locus 1 is functionally associated with the phenotype and nearby locus 2 is in LD with locus 1 with a  $r^2$  value of .33. In order to have the same power to detect association between the phenotype and locus 2, the sample size needs to be expanded with a factor 3 ( $= 1/r^2$ ).

Besides the poor interpretation of intermediate values,  $D'$  is biased upward inversely with sample size and  $r^2$  is typically lower at any distance. (Weiss and Clark, 2002).

In this study,  $r^2$  is used as the measure for LD rather than  $D'$ , because of the better interpretation of intermediate values and because of the tendency of  $D'$  to be overestimated with smaller sample sizes.

### *LD versus distance*

A commonly used measure for the extent of LD is its 'half-length', i.e. the distance at which the average LD drops below a certain threshold. For  $D'$ , this threshold typically is .5 (Reich et al., 2001). For  $r^2$ , the threshold depends on the sample size that can be analyzed, but a threshold of  $r^2$  equal to or greater than .33 is commonly considered to be useful (Ardlie et al., 2002). Although LD generally decreases with distance, a high portion of physically close marker pairs are not in strong LD with each other. Here, the extent of LD was defined as the distance over which the majority of marker pairs is useful for LD mapping, i.e. 50% or more of the marker pairs represent  $r^2$  values of 0.3 or higher.

In both the LW and LR panels,  $r^2$  values of .33 or higher occur in greater than 50% of the pairs at distances up to 40 kb. This indicates that at least half of the marker pairs at distances up to 40 kb are suitable for application in association studies, assuming that

the population sizes used can be up to 3 times as large as would be needed to perform the association study with the causative SNP or one in full LD herewith.

In both panels, pairs of markers separated by 60 to 80 kb seldom show useful values for LD (one of 35 for LW and 1 of 28 for LR). In both the LW and the LR panel, 5 pairs at distances of approximately 160 kb were found to represent useful LD, corresponding to 8-14 % of the marker pairs. For physically unlinked loci 1-2.5 % of marker pairs show useful pairwise LD values in all three population panels.

In conclusion, useful LD in the genomes of both these White breeds is estimated to extend for approximately 40 to 60 kb, although even a fraction of the distant marker pairs still represents useful LD.

In the M panel,  $r^2$  values of .33 or higher occur in only 18 of the 72 (25%) pairs at distances up to 40 kb. As a consequence, only a quarter of all pairs of markers would be useful in association studies. Only 4 (11%) of the pairs of markers 60 to 80 kb apart would be suitable for LD mapping. Although useful LD was detected between some pairs of markers approximately 160 kb apart, these only represents a small fraction of all pairs and in general useful LD over these distances seems to be rare. Overall, the  $r^2$  values observed in the M panel tend to be lower than observed in both the LW and LR panel. A good estimation of the extent of LD is complicated due to the fact that only a quarter of the pairs showing useful LD even at very short distances. The overall low LD values in the Meishan panel might be caused by the ascertainment bias (see below).

Comparing the fractions of pairs with useful LD at 0-40kb and 60-80kb, however, shows that the fraction has decreased from 25% to 11%. Although the fraction of marker pairs in useful LD is not above 50 % at any distance, the decrease of the portion of marker pairs in useful LD suggests that the extent of LD in Meishan is probably comparable to or somewhat larger than that in both White panels.

### *Long range LD*

To evaluate the decay of LD over longer distance, the 4 BACs that originate from the same contig were analysed. Distances between the BACs were inferred from the syntenic mouse sequence. In total the contig covers approximately 2500 kb of genomic sequence. BAC E is located near the left end of the contig. Considering BAC E as the starting point, BACs F, C and G are located 1200, 1500 and 2250 kb further downstream, respectively. This particular region contained over 500 marker pairs for each breed. For all three breeds, useful LD was observed at any distance up to 1800 kb for about 5 % of the marker pairs (Figure 2), which is higher than for unlinked loci (1-2.5%).

This absence of long-range LD contradicts data reported by by Nsengimana et al (2004), who demonstrated presence of LD useful for mapping purposes between marker pairs 3 to 10 cM apart, a distance corresponding to 3 to 10 Mb of genomic DNA. Probably, this

notable difference in the extent of LD is due to two factors. First, Nsengimana et al. used  $D'$  as the measure of LD and considered values of  $D' > 0.5$  to be useful. As was discussed above, values of  $D'$  are typically higher than values of  $r^2$ , and  $D'$  tends to be biased upwards with smaller sample sizes. The resulting overestimation of LD might result in a slower decrease of LD with distance. Second, in that study multiallelic microsatellite markers were used. Microsatellite markers have a higher mutation rate (and therefore have many newer alleles) rather than the older SNPs used in this study. Because LD around a newly established mutation tends to erode with time, the older haplotypes around SNP markers are expected to be shorter than those around relatively new microsatellite markers.

#### *Ascertainment bias.*

Ascertainment bias refers to the consequences of usage of SNPs that have lower minor allele frequencies in some lines. It is known that a reduction in allele frequencies leads to lower amounts of LD per chromosomal distance (Weiss and Clark, 2002). Because the SNPs were identified and selected from a panel representing all three breeds used in this study, the ascertainment bias is minimized. Nevertheless, for most SNPs the alleles are not equally distributed over the breeds (data not shown). For most markers, the minor allele frequency was highest in the LW and the LR panel. The overall low LD values observed in the Meishan panel might be influenced by the low minor allele frequencies in Meishan.

In conclusion, the extent of (useful) LD in both lines of White origin is estimated to be 40 to 60 kb. In Meishan, however, the extent could not be adequately estimated, due to low allele frequencies that led to overall low LD values.

The inverse relationship between LD and physical distance will make it possible to use LD for fine-mapping. Assuming that the results presented here are representative for the whole genome, a whole genome association scan would require genotyping of at least 60,000 SNPs per individual. This number, however, is an underestimation because more than one marker might be required to discriminate between regions with more than two haplotypes. The high number of markers required in a whole genome association study will probably make this approach too demanding and LD mapping or association studies to fine-map previously identified QTL regions will probably be more effective.



## References

- Aerts JA, Jungerius BJ, Groenen MGM (2004) POSA: Perl objects for sequence analysis. Submitted.
- Ardlie KG, Kruglyak L, Seielstad M. (2000) Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet.* 3(4):299-309.
- Botstein D, Risch N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet.* 33 Suppl:228-37.
- Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet.* 63(2):595-612.
- Devlin B, Risch N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics.* 29(2):311-22.
- Fahrenkrug SC, Freking BA, Smith TP, Rohrer GA, Keele JW. (2002) Single nucleotide polymorphism (SNP) discovery in porcine expressed genes. *Anim Genet.* 33(3):186-95.
- Goddard KA, Hopkins PJ, Hall JM, Witte JS. (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet.* 66(1):216-34.
- Janss LL, Van Arendonk JA, Brascamp EW. (1997) Segregation analyses for presence of major genes affecting growth, backfat, and litter size in Dutch Meishan crossbreds. *J Anim Sci.* 75(11):2864-76.
- Jungerius BJ, Rattink AP, Crooijmans RP, van der Poel JJ, van Oost BA, te Pas MF, Groenen MA. (2003) Development of a single nucleotide polymorphism map of porcine chromosome 2. *Anim Genet.* 34(6):429-37.
- Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK. (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am J Hum Genet.* 66(6):1882-99.
- Kruglyak L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet.* 22(2):139-44.

Moffatt MF, Traherne JA, Abecasis GR, Cookson WO. (2000) Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. *Hum Mol Genet.* 9(7):1011-9.

Nei M, Li WH. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A.* 76(10):5269-73.

Nezer C, Collette C, Moreau L, Brouwers B, Kim JJ, Giuffra E, Buys N, Andersson L, 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(1):277-85.

Pritchard JK, Przeworski M. (2001) Linkage disequilibrium in humans: models and data. *Am J Hum Genet.* 69(1):1-14.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. (2001) Linkage disequilibrium in the human genome. *Nature.* 411(6834):199-204.

Rieder MJ, Taylor SL, Clark AG, Nickerson DA. (1999) Sequence variation in the human angiotensin converting enzyme. *Nat Genet.* 22(1):59-62.

Risch NJ. (2000) Searching for genetic determinants in the new millennium. *Nature.* 405(6788):847-56.

Sobel E, Lange K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet.* 58(6):1323-37.

Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF. (2000) Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am J Hum Genet.* 66(1):69-83.

Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, Andersson G, Georges M, Andersson L. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature.* 425(6960):832-6.

Weiss KM, Clark AG. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18(1):19-24.

## **Chapter 7**

### **General Discussion**



The objective of the study described in this thesis was to explore the implementation and applications of single nucleotide polymorphisms (SNPs) in animal breeding and genetics. In this chapter the main findings of this thesis and the contribution to the field of animal breeding and genetics are discussed.

This chapter consists of three sections. In the first section the possibilities for SNP discovery (Chapters 2 and 3) and for SNP genotyping (Chapter 4) are discussed. In addition, a comparison of the general properties of SNPs and microsatellite markers is made. The second section addresses QTL mapping and the impact of identification of a single base mutation that influences lean muscle growth and backfat thickness (Chapter 5). The third section focuses on linkage disequilibrium (LD; Chapter 6) and addresses measures of LD and the extent of LD, together with its consequences for association studies.

## **Single Nucleotide Polymorphisms**

### *SNP discovery methods*

The need for *de novo* discovery of SNPs obviously decreases with the abundance of SNPs in public databases (for all species, 17 million in dbSNP alone, as of May 2004). Despite their abundance, however, selection of SNPs that are relevant for the projected research purpose remains a challenging task. While selection of SNPs in a specific genomic region might appear straightforward, selection of SNPs that are informative (i.e. polymorphic) in a specific (sub-) population will be more problematic.

Although the number of porcine SNPs in the databases has expanded over the last years (1545 SNPs, as of May 2004) only few porcine SNPs were publicly available at the start of this research project. Because additional markers were required for fine-mapping of the QTL for backfat thickness identified on chromosome 2 (SSC2; see below), this project aimed at *de novo* discovery of SNPs on this chromosome (as described in Chapter 2). To discover SNPs, several copies of each locus must be sampled from a population and compared for sequence differences. Below, two strategies to discover SNPs are discussed. The strategies mainly differ in the way the compared DNA sequence information has been generated or gathered.

### *Clone based resequencing*

Libraries of cloned DNA fragments have been generated for various purposes, including bacterial artificial chromosome (BAC) libraries representing an entire genome, plasmid libraries representing a (part of) the genome or expressed sequence tag (EST) libraries representing mRNA expressed in specific tissues. These libraries provide a convenient source of sequences for comparison to identify SNPs. Typically, sequencing of a library

of clones is done with a single universal vector-specific primer. Sequencing these clones has the advantage that they are always haploid: double peaks are always artefacts and the combination of alleles presented is a 'real' haplotype. The identified SNPs will be 'randomly' distributed over the genome, because there has been no *a-priori* focus on a region. Comparison of sequences from EST libraries can reveal coding region SNPs, which are particularly interesting because of their potential to change the functional properties of the encoded protein. It should be noted, however, that the polymorphism density in coding sequence typically is fourfold lower compared to random genomic sequence (Li and Sadler, 1991; Nickerson, 1998) and that most SNPs in coding regions represent silent mutations, i.e. they do not give rise to a change in amino acid sequence. Most libraries represent a maximum of two alternative copies of the genome, because they were generated from a single individual. Consequently, the number of SNPs identified from a single library will be low, but this will increase dramatically when comparing two or more libraries of different origin.

Finally, sequencing a library of clones to identify polymorphic positions is a major task, because the sequence reads should cover the library sufficiently to ensure overlap between the derived sequences. In a whole genome shotgun approach, it takes several-fold coverage of the genome before SNPs are discovered (Weber and Myers, 1997).

To limit the number of reads required, a Reduced Representation Shotgun Sequencing (RRSS) method has been developed (Altshuler et al., 2000). In this approach, DNA of a panel of individuals is mixed and fragmented using restriction endonucleases. Subsequently, fragments in a certain size range are purified and cloned into a vector system. As an example, Altshuler et al. (2000) described the digestion of the whole human genome by the restriction endonuclease *Bgl*III. Digestion with this enzyme is estimated to yield one million distinct fragments with an average length of 3100 bp. Of these fragments, 26000 are 500 to 600 bp in size, together representing 0.5% of the genome. After sequencing 52000 clones an average twofold coverage was reached and this lead to the identification of over 47000 human SNPs.

Overall, SNP identification by sequencing of a library of cloned fragments requires a substantial sequencing effort to ensure SNP identification. Although the sequencing effort required can be decreased through RRSS, these approaches are feasible only in large-scale SNP discovery projects. The required sequencing effort can be largely reduced if a finished genome sequence of the species is available. Especially in case the sequenced library has a different origin than the library used in the genome sequencing project, all sequence reads generated can be readily compared to the genome sequence to identify polymorphisms and the library only needs to be skimmed at a low coverage. At the Beijing Genome Institute, this approach has been applied to identify 2.8 million SNPs by resequencing at 0.3x coverage from three libraries from different breeds (Liu et al., 2004). For pigs, A Danish-Chinese consortium aims at

obtaining sequence information of all pig genes and to work towards determination of the major part of the genome sequence. Several cDNA libraries are being sequenced and were screened for SNPs (<http://www.piggenome.dk>). At the time of writing, however, this effort is still ongoing and only little information is publically available yet.

### *Resequencing of PCR products*

Direct (re)sequencing of PCR products obtained from different individuals probably is the most straightforward SNP discovery method. This method requires DNA sequence information to facilitate PCR primer development. Because in Chapter 2 the aim was development of SNPs on SSC2, a variety of sequences that were mapped to SSC2 in other projects were used as a template for primer development. These sequence data originated from sample sequencing of SSC2 derived BAC clones (Rattink et al., 2001b) and from data of genes mapped to SSC2 retrieved from the Genbank database. In addition, some PCR products developed for mapping on a radiation hybrid panel (Rattink et al, 2001a) were also resequenced. Because PCR products known to originate from SSC2 were resequenced, all identified polymorphisms are expected to map to this chromosome. Using existing sequence information, however, will not identify SNPs distributed equally over the genome, but will lead to an overrepresentation of regions that were studied previously. In Chapter 2, all PCR products were resequenced from a panel of eight individuals, each representing a different founder line of the Wageningen QTL mapping population. The size of this panel assures a sufficient coverage of the sequenced regions, whereas the presence of the founder lines in the panel helps identification of SNPs that are of use in fine-mapping in the QTL mapping population. Because of the dependence on PCR primer development, the resequencing method will become too inefficient in large-scale SNP discovery projects. In addition, the common usage of genomic DNA as a template results in diploid sequence reads, which troubles the identification of read artefacts from heterozygous samples.

All together, resequencing of PCR products is a very efficient way to identify SNPs in a specific target region (i.e. on a specific chromosome, in a gene, exon or intron etc.). The method is very efficient to identify SNPs that are segregating in specific populations (i.e. by resequencing the founders of that population).

Summarising, the single best method to select or identify SNPs does not exist, as it is influenced by several factors. These factors include the availability of SNPs from databases, availability of a finished genome sequence, the available sequencing capacity, the choice of target region(s) and the population studied. Presently, the porcine genome has not yet been sequenced and the sequence capacity available within this project was not sufficient for a large-scale effort for constructing and sequencing a library with sufficient coverage.

Considering the availability of DNA sequence data for primer design and the aim to target SSC2 specifically, the resequencing of PCR products was an appropriate method for the work described in Chapter 2. In addition, the composition of the discovery panel assured identification of SNPs that would be relevant for fine-mapping in the Wageningen QTL mapping resource population.

### *SNP density and genome diversity*

In publications on SNP discovery, the average SNP spacing or the average SNP density is often reported as the number of bases per SNP (e.g. 500 bp/SNP) or as the number of SNPs per base (e.g. 1/500 or 0.002 SNP/bp). However, a direct comparison of SNP densities over different studies is only valid if both studies compared identical numbers of chromosomes or individuals for SNP discovery. As the number of polymorphisms to be observed is strongly dependent on the number of chromosomes sampled, the reported averages should be normalised for the assayed sample size. Nei and Li (1979) introduced an index termed nucleotide diversity or heterozygosity. This value reflects the rate of nucleotide difference between two randomly chosen chromosomes. In Chapter 2 of this thesis, a discovery panel of 8 individuals (thus 16 chromosomes) was used. Therefore the overall polymorphism rate of one SNP in every 108 bp corresponds to a heterozygosity index 1/358 (one SNP in every 358 bp). This indicates that the heterozygosity index in the porcine genome is comparable to that in cattle (1/443; Heaton *et al.*, 2001) or in mouse (Lindblad-Toh *et al.*, 2000). Heterozygosity in the human genome, however, is lower (1/1331; Sachidanandam *et al.*, 2001). The lower heterozygosity in human might reflect the reduced diversity, probably due to recent population expansion (Kaessmann *et al.*, 2001).

### *SNP as markers*

The predominant marker type in most mapping projects is the microsatellite marker. Microsatellites are commonly used because of their high number of alleles per locus. Consequently, microsatellites have a high information content and thus enabling a reduction in the number of loci to genotype. The alleles of the microsatellite are integer values, reflecting the sizes of the PCR products. Because the genotyping and analysis methods might influence the integer values assigned, it is difficult to compare data generated on different machines using different methods in different laboratories.

Over the last few years, SNPs gained in popularity as markers. SNP alleles are generally binary: either the one or the other allele is present. This has the advantage that allele calls are independent of the genotyping method or the laboratory and thus can be compared over studies.

Allelic variation in microsatellite markers is subject to recurrent mutations. Therefore microsatellites loci on different chromosomes have a reasonable chance of being



identical-by-state (IBS) without sharing a common ancestor. The mutation rate of SNP markers, however, is very low and this limits the chances of a recurrent mutation to a level that can be ignored. As a consequence, SNP markers that are IBS are very likely to share at least one common ancestor and therefore are also identical by descent (IBD). This makes SNP markers very suitable for the detection of historical recombination events in long-term linkage studies and LD mapping.

The major disadvantage of SNPs is that they are bi-allelic and therefore are less informative than that of multi-allele microsatellite markers. In general, it has been estimated that linkage studies require approximately 2.5 SNP to provide the amount of information equal to that of one microsatellite marker (Kruglyak, 1997). This increase of markers to be genotyped, however, is compensated by today's SNP genotyping methods that facilitate high-throughput genotyping with highly automated data analysis.

### *SNP genotyping methods*

Over the last years, a wide variety of SNP genotyping methods has been reported and several reviews on this topic have been published (Syvänen, 2001; Kristensen et al., 2001).

Despite the variety in methods, most genotyping assays basically rely on a handful of principles: digestion with restriction enzymes, allele-specific oligonucleotide hybridisation (ASO), single base extension (SBE; or minisequencing), allele-specific primer extension, allele specific ligation or allele-specific amplification. For most of these reaction principles a variety of reaction formats (single wells, 96 or 384 well formats and multiplexes) and detection methods is available. Together, the combinations account for the numerous methods that are currently available.

Identification of a single base difference in the approximately 6 billion basepairs of a diploid mammalian genome is a demanding task (Syvänen, 2001). Consequently, most (if not all) genotyping methods require a PCR amplification to multiply the target region of the genomic region around the target SNP before the actual genotyping. Fortunately, the multiplex PCR amplification reduces the number of separate reactions. In a heterozygous sample, however, the alleles might not be amplified at identical rates and the fractions of both alleles might deviate from a 50/50 ratio. For PCR products that severely deviate from 50/50, discrimination between weak but true heterozygous signals from slightly elevated background noise will become difficult. Although multiplex PCR might be more prone to this phenomenon, it is also observed in singleplex PCR reactions.

With the wide variety of options for SNP genotyping, choosing the appropriate method can be a major task and it is impossible to give a general advice or ranking, because the final decision is influenced by several factors. First, the number of genotypes to be assayed should justify the effort required to develop the assay. Next, the method of

choice may be different for typing many SNPs from only few samples or only few SNPs from many samples. Then, the spacing of the SNPs should be considered: it matters whether the SNPs are close together in one region or even covered by one PCR product or whether they are scattered across the genome. Finally, availability of equipment and commercial kits should be taken in to consideration, as this will help to decrease the investments of money for machinery and of time for assay development.

In this project (Chapter 6), genotyping was performed by single base extension (SBE) and by resequencing. The choice for SBE was based on the availability of capillary sequencers (ABI 3100) and commercial kits (ABI SNaPshot Multiplex kit), both allowing multiplex template amplification and facilitating multiplex SBE reactions.

In addition, some SNPs were genotyped through resequencing, because several SNPs were close enough to be covered in a single sequencing read. The *asGenotypes* function of POSA (Chapter 3) was used to automate the analysis of the sequence reads generated for genotyping.

### **QTL fine-mapping**

#### *QTL mapping and the candidate gene approach*

Basically two approaches can be applied to identify genes involved in a trait: QTL analysis and candidate gene approach. In QTL mapping, markers are used to perform a whole genome scan and statistical analysis are applied to provide evidence that a certain genomic region is segregating together with a specific phenotype. In this approach, there is no need for *a priori* knowledge about the genetics, biochemistry or physiology of either the phenotype. A detailed marker map, however, is required. QTL mapping is especially suitable for detecting genes with a large effect, since it is difficult to reach significance levels for genes that only modestly contribute to the phenotype. Typically, QTL regions are still very large (tens of cM) and further fine-mapping is required to identify the gene (or mutation) underlying this QTL.

In the candidate gene approach the focus is on mutations in genes with an expected impact on the studied phenotype to detect association between the gene and the phenotype. For example, an association between a missense mutation in MC4R in a large number of individual animals from several different pig lines with backfat, growth rate, and feed in a number of lines intake was reported (Kim et al, 2000).

In practice, often a combination of both strategies is used: QTL mapping will reveal chromosomal regions that have an impact on the trait and subsequently functional candidate genes can be selected from this region. Regardless of the approach, identification of the causative mutation in a QTL region remains a challenging task. Only few cases have been reported in which a QTL region has been narrowed down to the underlying causative mutation. For example, Van Laere et al. (2003) described the

identification of the mutation that underlies the maternally imprinted QTL for backfat thickness. This QTL was identified in crosses between Large White (LW) X Pietrain (Nezer et al., 1999), LW X Wild Boar (Jeon et al., 1999) and European Whites X Meishan (Rattink et al., 2000; de Koning et al., 2000). Although the best positions of the QTL varied between the populations, the confidence intervals greatly overlapped and all three covered the IGF2 locus. IGF2 was the prime candidate gene to underlie the QTL for BFT on SSC2p, because of its position on the distal tip of SSC2p, its function in regulation of the trait and its mode of expression. A LD-based haplotype sharing approach identified a 250 kb haplotype including the *IGF2* gene and that is shared by both the LW X Pietrain and the LW X Wild Boar population. Therefore, this haplotype was predicted to carry the quantitative trait nucleotide (QTN, Mackay, 2001; Nezer et al., 2003). This region is predicted to harbour at least 12 genes of which eight are expressed in a Mendelian mode, two are expressed exclusively from the maternal allele and two are expressed exclusively from the paternal allele (INS and IGF2) (Nezer et al., 2003). An effort for resequencing of a 28.6 kb segment containing INS and IGF2 identified a shared haplotype with a total of 258 SNPs. Finally, the G to A transition at position 3072 in the third intron of IGF2 was shown to be the QTN underlying the QTL in both the LW X Pietrain and the LW X Wild Boar population.

In Chapter 5, the backfat thickness QTL in the Wageningen Meishan X White cross is re-analyzed with respect to the IGF2-intron3-G3072A mutation. At first, the QTL in the Wageningen population was expected to be different from the QTL in both other populations because the initial imprinting analysis (following the model proposed by Knott et al., 1998) revealed a maternally imprinted QTL on SSC2 at 63 cM. Although the QTL covered over 60 cM, no significance levels were reached near the IGF2 locus. (Rattink et al., 2000). Imprinting analysis of the same data (following the imprinting model described by de Koning et al., 2000) showed that the best position of the paternally expressed QTL was located at 36 cM. Although the QTL exceeded the significance threshold around the IGF2 locus, the presence of a second, more proximal QTL could not be excluded.

The identification of the causative mutation in the other two populations made it possible to test the Wageningen population for this mutation and incorporate this information in a re-analysis of the QTL. As is reported in Chapter 5, the causative mutation segregated within the Wageningen population and it is shown that the mutation explains most of the QTL effect observed on this chromosome. In conclusion, the effect of the IGF2-intron3-G3072A mutation on muscle growth and backfat thickness reported by Van Laere et al. (2003) was verified to also cause the main QTL effect in the Wageningen population.

Summarizing, the observed maternally imprinted QTLs in all three populations were caused by the same mutation, despite the former differences in both best position of the QTL and pattern of the QTL graph. This suggests that it might be more efficient to describe a QTL by giving a confidence interval rather than a best position. This will facilitate a better comparison of QTL over populations and studies.

In most QTL mapping populations, the number of informative meioses in the population will be a limiting factor in fine-mapping. In a LD-based haplotype sharing analysis of consensus QTL regions over populations, not only meioses in the experimental populations will provide useful information, but also information about historical recombination's is provided that will contribute to the identification of haplotypes associated with the phenotype.

### *Implications of the causative mutation*

#### *QTN satisfies scientists, not breeders*

Without knowledge on the causative mutation, animal breeders will still be able to apply marker assisted selection (MAS) to select for the favourable QTL allele. The association between a linked marker and the causative mutation, however, might be disrupted by a recombination event and therefore association between the marker and the phenotype has to be monitored in new generations. Although a closely linked, highly correlated marker for a trait is sufficient to implement is the marker in breeding programs, the identification of the causative mutation will still be the goal from a scientific point of view.

#### *One QTN does not explain a complex trait*

As was described above, a mutation with a large effect on backfat thickness has been identified and the resolution of the imprinted QTL on SSC2p is reduced to a single base pair. The identification of this causative mutation, however, only means that now this QTL is resolved. A complex trait like BFT, however, is influenced by many genetical and environmental factors and it will still be a long way towards understanding the genomic architecture of BFT. Identification of the IGF2 mutation, however, will contribute in a better understanding of the trait and will give insight in the involved pathways. Moreover, it is a relevant factor to test for in breeding programs. In experimental populations, all assessable factors known to influence the phenotype can be used to enhance the model and thus will enable a better evaluation of other, remaining factors involved, as was done in Chapter 5.

## **Linkage Disequilibrium**

### *Measures of LD*

As was discussed in Chapter 1, several measures of LD exist, of which  $D'$  and  $r^2$  are most commonly used. In Chapter 6 of this thesis,  $r^2$  was used because of the better

interpretation of intermediate values and because it is overestimated with small sample sizes. Although  $r^2$  has higher practical relevance, it should be noted that the calculation of  $r^2$  is only possible for pairs of biallelic markers (like most SNPs). Calculating LD between microsatellite markers, which typically have more than two alleles, might be problematic and  $r^2$  might thus not be applicable for this type of markers.

Intermediate values of  $r^2$  can be easily interpreted in terms of sample size needed to detect association. Consider a two locus model in which the first locus is functionally associated with a phenotype and the second locus is in LD with the first locus. To have the same power to detect the association between the phenotype and the second locus as between the phenotype and the first, causative locus, the sample size must be increased with a factor  $1/r^2$ .

In association studies, the maximum available sample size is often limited. Assuming that the sample size in a study can be increased threefold,  $r^2$  values of  $1/3$  or greater reflect *useful* LD, i.e. LD useful for association mapping. When studying a rare human disease, the number of affected individuals might be limited and a three-fold increase in sample size might be problematic. Although practical and financial issues will become limiting, in animal breeding the available sample size can be increased by generation of experimental populations. Consequently, the threshold for useful LD is rather arbitrary and might be influenced by availability of resources. In Chapter 6, however,  $r^2$  values exceeding the threshold of  $1/3$  are considered as useful LD.

### *The extent of LD*

Typically, LD decreases with physical distance. A common way to define the relation between the relation between LD and physical distance is the 'half-length' of LD, i.e. the distance at which the value of  $D'$  drops below 0.5. With  $r^2$ , the extent of useful LD can be defined as the distance over which  $r^2$  drops below  $1/3$ . The measure used has a notable influence on the estimation of the extent of LD, as was illustrated by Weiss and Clark (2002), who re-analyzed the data previously reported by Reich et al. (2001). Reich et al. (2001) used  $D'$  and estimated the extent of LD in a North American population of European decent to be approximately 60 kb. Weiss and Clark (2002) analysed the same data on the same population, but the  $r^2$  value dropped below  $1/3$  at approximately 30 kb. This clearly demonstrates that the extent over which LD is useful for mapping purposes is overestimated when using the half-length of  $D'$ .

Despite the strong relation between LD and distance, a large fraction of physically close marker pairs does not show strong LD. In Chapter 6, the extent of LD was determined as the distance over which more than 50% of the marker pairs still show useful LD. This alternative approach has the advantage that it is based on the fraction of marker pairs that are in useful LD, rather than reflecting the mean value of LD for all pairs to drop below a certain threshold. Because this alternative definition directly reflects the fractions

of marker pairs in useful LD, it might serve as an indicator for average marker densities needed in association mapping studies.

In Chapter 6, the extent of LD in the porcine genome was estimated through analysis of SNP markers in seven genomic regions. Although in some cases LD may extend over longer distances, useful LD ( $r^2 \geq 1/3$ ) occurs for 50% of marker pairs at distances of 40 to 60 kb in two Dutch White commercial lines of predominant Large White or Dutch Landrace origin.

This extent of useful LD in the porcine genome is of the same order of magnitude as the extent of LD reported in human: Abecasis et al. (2001) reported that useful LD was observed for 50 % of marker pairs at distances up to 50kb.

Previously, LD was expected to extend further in livestock genomes than in the human genome due to intensive artificial selection accompanied by a reduced effective population size (Haley, 1999). A comparison between LD in the two genomes, however, might not be valid because of the differences in population characteristics and population dynamics, which both influence LD and its extent.

Until recently, only little information on the extent of LD in the porcine genome has been published. Nsengimana et al. (2004) analyzed pairwise LD between microsatellite markers from two genomic regions and reported that LD extends over 3 to 10 cM (corresponding to roughly 3 to 10 Mb). The finding of LD among microsatellites that are separated by several centi-Morgans is not consistent with our data derived with SNPs (Chapter 6). Although the explanation for this discrepancy is (yet) unclear, it might involve the greater power to detect LD between multi-allelic markers and the difference in the measures of LD ( $D'$  vs.  $r^2$ ). In addition, microsatellite markers have a higher mutation rate (and therefore have many newer alleles) rather than the older SNPs used in our study. Because LD around a newly established mutation tends to erode with time, the older haplotypes around SNP markers are expected to be shorter than those around relatively new microsatellite markers.

### *LD and association mapping*

Several publications estimated the average extent of LD in the human genome by studying the decay of LD over distance in a number of genomic regions (Dunning et al., 2000; Reich et al., 2001; Abecasis et al., 2001). It should be noted, however, that the average extent of LD is not a good guide for the design of LD mapping approaches, because of the tremendous variability in the extent of LD from one region to another, and because some markers do not show sufficient LD with their environment, even within regions of high LD. Consequently, the extent of LD in the region of interest should be empirically assessed in order to determine the appropriate density of markers for fine-mapping. To characterize the local patterns of LD in the human genome an effort to

construct a haplotype map of the human genome (named the HapMap Project) was initiated by a large international consortium (The HapMap Consortium, 2003).

The haplotype-block model provides a simple design for LD-based mapping studies. In outline, the main haplotypes in each haplotype block are identified, followed by the determination of a minimal set of SNPs needed to distinguish among these haplotypes (often referred to as haplotype tagging SNPs). This set of tagging SNPs is then used to identify haplotype blocks that show association with the studied phenotype. In association mapping, the size of haplotype blocks has both favourable and unfavourable consequences. In case the haplotype blocks are large, fewer haplotype tagging SNPs need to be genotyped in a whole genome scan and to narrow down the region to a single block. Identification of the polymorphism that actually causes a difference in the phenotype, however, will be a major task because the large block will contain more genes and more polymorphisms, most of which are in LD. Small haplotype blocks, on the other hand, require genotyping of more tagging SNPs to identify the block of interest, but because it is shorter, it will contain less candidate polymorphisms to be validated.

It has been estimated that, in human non-African populations, 300000 SNPs are needed for this approach (Gabriel et al., 2002). Although this procedure seems simple, problems may arise as every block is treated as completely independent, while substantial, though not complete, relations may exist between one block and another. As a result, a unique assignment of polymorphic sites to blocks might not be possible, and even definitions with well-defined criteria do not facilitate a unique assignment of sites to blocks.

### *Effect of selection on LD*

In Chapter 6, the average extent of LD is reported to be approximately 40 kb in two Dutch commercial lines. The mutant allele in the IGF2 gene (analysed in Chapter 5) promotes lean muscle growth, individuals carrying this allele are likely to be selected to contribute to the next generation and thus a larger than expected fraction of the population will bear the haplotype carrying the beneficial mutation. Consequently, the haplotype containing the mutation will contribute more to next generations and as a result, this haplotype will be observed more often than expected. In case the mutation is relatively recent, the haplotype has not yet eroded in time and long haplotype blocks will be observed around beneficial alleles.

As mentioned in Chapter 5, the regulatory mutation in the IGF2 gene is part of a 250 kb haplotype, which is considerably longer than the 40- to 60 kb average extent of LD estimated in Chapter 6. The presence of long haplotypic structures might suggest the presence of recent positive selection. Similarly, extraordinary long extents of LD were detected around certain alleles that contribute to human malaria resistance (Sabeti et al., 2002). These examples suggest that regions that underwent selection pressure will contain larger haplotype blocks. This, however, might only hold for haplotypes around

beneficial alleles. Through selection beneficial alleles contribute more than expected to a next generation, this is sometimes referred to as a selective sweep. In case of a new mutation with a negative effect (e.g. a mutation that contributes to disease risk), there is not a single beneficial haplotype. Therefore, mutations contributing to disease risk will not undergo a selective sweep. Consequently, these mutations with a negative effect will not be part of larger haplotypic structures and will be harder to identify.

The aim of the work described in this thesis was the implementation and application of SNP markers in animal breeding and genetics. The emphasis was on the analysis of fatness traits in pigs, in particular of the imprinted QTL region on SSC2p. The identification of SNP markers in this region is described in Chapter 2. Chapters 3 and 4 describe POSA and SNPTyper, both are software tools that contribute to automation of data analysis in SNP discovery and SNP genotyping. The QTL on SSC2p was reanalysed after the identification of a SNP in the IGF2 gene that had high impact on lean muscle growth and back fat thickness. In Chapter 5 this re-analysis is described and it is concluded that the QTL is reduced to a quantitative trait nucleotide or QTN. For research in other genomic regions, the SNP marker density needed in association studies was an issue of discussion. To address this question, Chapter 6 describes an estimation of the extent of LD in the porcine genome. Although an average extent of LD was estimated, further investigation of LD and its extent is crucial in order to develop a map of the patterns of sequence variation in the porcine genome.

## References

- Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WO. (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet.* 68(1):191-197.
- Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407(6803):513-6.
- de Koning DJ, Rattink AP, Harlizius B, van Arendonk JA, Brascamp EW, Groenen MA. (2000) Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proc Natl Acad Sci U S A.* 97(14):7947-50.
- Dunning AM, Durocher F, Healey CS, Teare MD, McBride SE, Carlomagno F, Xu CF, Dawson E, Rhodes S, Ueda S, Lai E, Luben RN, Van Rensburg EJ, Mannermaa A, Kataja V, Rennart G, Dunham I, Purvis I, Easton D, Ponder BA. (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet.* 67(6):1544-54.



Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. (2002) The structure of haplotype blocks in the human genome. *Science* 296(5576):2225-9.

Haley, CS (1999) Advances in QTL mapping, pp 47-59 in *Proceedings of From Lush to Genomics: Visions for Animal Breeding and Genetics*, edited by Dekkers, JCM, Lamont, SJ and Rothschild M. Iowa State University, Ames, Iowa, USA.

Heaton MP, Grosse WM, Kappes SM, Keele JW, Chitko-McKown CG, Cundiff LV, Braun A, Little DP, Laegreid WW. (2001) Estimation of DNA sequence diversity in bovine cytokine genes. *Mamm Genome*. 12(1):32-7.

Jeon JT, Carlborg O, Tornsten A, Giuffra E, Amarger V, Chardon P, Andersson-Eklund L, Andersson K, Hansson I, Lundstrom K, Andersson L. (1999) A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nat Genet*. 21(2):157-8.

Kaessmann H, Wiebe V, Weiss G, Paabo S. (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat Genet*. 27(2):155-6.

Kim KS, Larsen N, Short T, Plastow G, Rothschild MF. (2000) A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. *Mamm Genome* 11(2):131-5.

Kristensen VN, Kelefiotis D, Kristensen T, Borresen-Dale AL. (2001) High-throughput methods for detection of genetic variation. *Biotechniques*. 30(2):318-326.

Kruglyak L, Nickerson DA. (2001) Variation is the spice of life. *Nat Genet*. 27(3):234-6.

Kruglyak L. (1997) The use of a genetic map of biallelic markers in linkage studies. *Nat Genet*. 17(1):21-4.

Li WH, Sadler LA. (1991) Low nucleotide diversity in man. *Genetics* 129(2):513-23.

Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette JP, Ardlie K, Reich DE, Robinson E, Sklar P, Shah N, Thomas D, Fan JB, Gingeras T, Warrington J, Patil N, Hudson TJ, Lander ES. (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet*. 24(4):381-6.

Liu B, et al. (2004) A polymorphism map of chicken with 2.8 million SNPs. Submitted.

Mackay TF. (2001) The genetic architecture of quantitative traits. *Annu Rev Genet.* 35:303-39.

Nei M, Li WH. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A.* 76(10):5269-73.

Nezer C, Collette C, Moreau L, Brouwers B, Kim JJ, Giuffra E, Buys N, Andersson L, 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(1):277-85.

Nezer C, Moreau L, Brouwers B, Coppieters W, Dettloux J, Hanset R, Karim L, Kvasz A, Leroy P, 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(2):155-6.

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet.* 19(3):233-40.

Nsengimana J, Baret P, Haley CS, Visscher PM. (2004) Linkage disequilibrium in the domesticated pig. *Genetics.* 166(3):1395-404.

Rattink AP, Jungerius BJ, Faivre M, Chardon P, Harlizius B, Groenen MA. (2001a) Improving the comparative map of SSC2p-q13 by sample sequencing of BAC clones. *Anim Genet.* 32(5):274-80.

Rattink AP, Faivre M, Jungerius BJ, Groenen MA, Harlizius B. (2001b) A high-resolution comparative RH map of porcine chromosome (SSC) 2. *Mamm Genome.* 12(5):366-70.

Rattink AP, De Koning DJ, Faivre M, Harlizius B, van Arendonk JA, Groenen MA. (2000) Fine mapping and imprinting analysis for fatness trait QTLs in pigs. *Mamm Genome.* 11(8):656-61.

Reich DE, Gabriel SB, Altshuler D. (2003) Quality and completeness of SNP databases. *Nat Genet.* 33(4):457-8.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF,

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D;

International SNP Map Working Group. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. 409(6822):928-33.

Ward R, Lander ES. (2001) Linkage disequilibrium in the human genome. *Nature*. 411(6834):199-204.



## Summary



Over the last decades, there has been an enormous increase in knowledge of the genomes of livestock species. Genetic maps of livestock genomes have been applied in several linkage studies to map loci and genes that underlie genetic variance of economically important traits. In pig production, fatness is an important trait that has been extensively studied over the last years and several quantitative trait loci (QTL) for fatness traits have been identified. Most QTL regions, however are still large and further fine-mapping is required to identify the underlying gene or mutation.

Several studies reported a quantitative trait loci (QTL) for backfat thickness (BFT) on the p-arm of porcine chromosome 2 (SSC2p), but the identified region still spans a large part of the chromosome. Additional markers were required to fine-mapping of this region. In **Chapter 2**, the development of single nucleotide polymorphism (SNP) markers within the QTL region is described. The SNPs were identified by comparative resequencing of polymerase chain reaction (PCR) products from a panel of eight individuals. The panel consisted of five Large Whites (each from a different Dutch breeding company), a Meishan, a Pietrain and a Wild Boar. In total, 67 different PCR products were sequenced and 301 SNPs were identified in 32429 bp of consensus sequence, an average of one SNP in every 108 bp. After correction for sample size, this polymorphism rate corresponds to a heterozygosity value of one SNP in every 357 bp. For 63% of the SNPs, there was variation among the five Large Whites, and these SNPs are relevant for linkage and association studies in commercial populations. Comparing the Whites with other breeds revealed higher variation rates with: (i) Meishan, 89%; (ii) Pietrain, 69%; (iii) Wild Boar, 70%. Because many of the experimental populations to identify QTL are based on crosses between these breeds, these SNPs are relevant for the fine mapping of the QTL identified within these crosses.

Identification of SNPs is often done by analyzing and comparing DNA sequences. This analysis can be a laborious and time-consuming task, but the time needed can be dramatically reduced by implementation of software for automated data processing. Most available software solutions, however, still require user intervention or provide modules that need advanced informatics skills to allow implementation in pipelines. In **Chapter 3**, POSA is presented: a pair of new perl objects that describe DNA sequence traces and phrap contig assemblies in detail. These objects allow a flexible and easy design, implementation and usage of perl-based pipelines, while requiring only little programming skills. Features include basecalling with quality scores (Phred), contig assembly (Phrap), generation of primer3 input and automated SNP annotation (PolyPhred). Using POSA, analysis time needed for annotation of SNPs in a contig of resequenced PCR products can be reduced from several hours to few minutes.

SNPs are mostly used as genetic markers and therefore the number of SNP genotypings performed is increasing. Although a wide variety of SNP-genotyping techniques has been described, most techniques still have low throughput or require major investments. For laboratories that have access to an automated sequencer, a single-base extension (SBE) assay can be implemented using the ABI SNaPshot™ kit. In **Chapter 4**, we present a modified protocol comprising multiplex template generation, multiplex SBE reaction, and multiplex sample analysis on a gel-based sequencer such as the ABI 377. These sequencers run on a Macintosh platform, but on this platform the software available for analysis of data from the ABI 377 has limitations. First, analysis of the size standard included with the kit is not facilitated. Therefore a new size standard was designed. Second, using Genotyper (ABI), the analysis of the data is very tedious and time consuming. To enable automated batch analysis of 96 samples, with 10 SNPs each, we developed SNPtyper. This is a spreadsheet-based tool that uses the data from Genotyper and offers the user a convenient interface to set parameters required for correct allele calling. In conclusion, the method described will enable any lab having access to an ABI sequencer to genotype up to 1000 SNPs per day for a single experimenter, without investing in new equipment.

Previously, several experimental crosses in pigs had been used to detect QTLs for a variety of production traits like fatness and meat quality. A paternally expressed QTL for backfat thickness (BFT) was identified near the *IGF2* locus on the distal tip of SSC2p in three experimental F2-populations. Recently, a mutation in a regulatory element of the *IGF2* gene was identified as the Quantitative Trait Nucleotide (QTN) underlying the major QTL effect on muscle growth and BFT in crosses between Large White and Wild Boar or Pietrain. In **Chapter 5**, it is demonstrated that the *IGF2* mutation also controls the paternally expressed QTL for backfat thickness in a cross between Meishan X European Whites. In addition, a comparison of QTL for backfat thickness measured by Hennessy grading probe and by ultrasound measurement (USM) revealed remarkable differences. In the USM analyses, the *IGF2* mutation explains the entire QTL effect on SSC2p, whereas in the BFT-HGP analysis the presence of a second minor QTL can not be excluded. Finally, **Chapter 5** shows that the *IGF2* mutation has no pleiotropic effect on a paternally expressed QTL for teat number identified on SSC2 in the same population.

As was mentioned before, many studies were performed to quantitative trait loci (QTL) that affect traits with commercial relevance in livestock species. Typically, the resolution of QTL mapping is not sufficient to identify the genes or mutations that underly the QTL. A alternative approach to identify mutations that influence the phenotype is to search for an association between a specific variant (allele) and the phenotype, by comparing



groups of extremes for the trait, like affected and unaffected individuals. In a region, however, SNPs can be part of specific haplotype structures and these are therefore in linkage disequilibrium (LD). Linkage disequilibrium refers to the correlation among neighboring alleles, reflecting non-random patterns of association between alleles at (nearby) loci. A better understanding of LD in the porcine genome is of direct relevance for identification of genes and mutations with a certain effect on the traits of interest. In **Chapter 6**, SNP markers in four genomic regions are analyzed in three panels from populations of Large White, Dutch Landrace and Meishan origin. Based on these data, it is estimated that useful LD extents over approximately 50 kb in the porcine genome.

In **Chapter 7**, the main findings of described in Chapters 2 through 6 are discussed in three sections. In the first section the possibilities for discovery and for genotyping of single nucleotide polymorphisms (SNPs) are discussed. In addition, a comparison with of the general properties of SNPs and microsatellite markers is made. The second section addresses QTL mapping and the impact of identification of the IGF2 mutation that influences lean muscle growth and backfat thickness. The third section focusses on linkage disequilibrium (LD) and discusses measures of LD and the extent of LD, together with their consequences for the design of association studies.

The aim of this thesis was the implementation and application of SNP markers in animal breeding and genetics. The emphasis was on the analysis of fatness traits in pigs, in particular of the imprinted QTL region on SSC2p. Several aspects of the implemetation of SNP markers in genetic analysis of livestock were addressed, including SNP discovery methods and tools for enhancing sequencing and genotyping data analysis. The QTL on SSC2p was reanalysed after the identification of a SNP in the IGF2 gene that had high impact on lean muscle growth and back fat thickness. This re-analysis showed that the QTL region is narrowed down to a single base pair: the quantitative trait nucleotide (QTN). For research in other genomic regions, the SNP marker density needed in association studies was an issue of discussion. To address this question, the extent of LD in the porcine genome was studied. It was estimated that the majority of marker pairs separated by less than 50 kb show levels of LD that might be useful in association studies. Further investigation of LD and the extent of LD, however, is crucial in order to develop a map of the patterns of sequence variation in the porcine genome.



## **Samenvatting**



Erfelijke kenmerken liggen gecodeerd in de genen. Genen maken (een klein) deel uit van het genoom dat bij zoogdieren bestaat uit ongeveer 3 miljard baseparen. Iedere cel van een individu bevat 2 kopieën van het genoom en dus circa 6 miljard baseparen. Er zijn vier verschillende basen (A, C, G en T) en de volgorde hiervan wordt wel de sequentie genoemd. Het merendeel van deze sequentie is nagenoeg identiek tussen individuen. Voor een aantal soorten, waaronder mens, muis, hond en kip, is de totale genoomsequentie bekend. In het in dit proefschrift besproken onderzoek was aan het varken gewijd. De genoomsequentie van het varken is dit nog niet bekend, maar dit zal waarschijnlijk binnen enkele jaren wel het geval zijn.

Het DNA van ieder individu is voor de helft van de vader en voor de helft van de moeder afkomstig. In genetisch onderzoek is het van belang te weten welke helft van welke ouder komt. Om dit te bepalen worden stukjes DNA gebruikt die kleine verschillen vertonen tussen de ene en de andere kopie van het genoom. Deze stukjes DNA noemt men merkers en de verschillende varianten van deze merkers heten allelen.

Een veelgebruikt type merker is de microsatelliet. Deze bestaat uit een stukje specifieke sequentie met daarin een herhalend stukje (bv. CT of GA) waarbij de allelen bestaan uit verschillen in het aantal herhalingen (en dus de lengte) van dit stukje sequentie.

In dit proefschrift wordt echter vooral aandacht besteed aan een ander type merker: single nucleotide polymorphisms (SNPs). De SNP wordt zo genoemd omdat het meestal gaat omdat een base (nucleotide) verschillend is, terwijl het merendeel van de omliggende sequentie hetzelfde is. Gemiddeld komt een SNP eens in de paar honderd basen voor en er zijn er dus miljoenen SNPs in het genoom. De meeste SNPs kennen twee allelen, waarbij de meest voorkomende combinaties A of G en C of T zijn.

In Hoofdstuk 2 wordt beschreven hoe naar SNPs werd gezocht in een bepaald gebied van het genoom, namelijk de korte arm van varkenschromosoom 2.

In grote lijn komt het erop neer dat stukjes DNA gekopieerd worden door middel van PCR; een techniek die specifieke stukjes DNA met lengtes van ongeveer honderd tot enkele duizenden basen kan vermenigvuldigen tot tienduizenden kopieën. De specificiteit zit in het ontwerp van de primers, korte stukjes DNA van 20-25 basen, die bepalen welk stuk van het originele genomisch DNA gekopieerd moet worden. De kopiën kunnen vervolgens gebruikt worden om de volgorde van de basen bepalen (sequenzen). Door dit te doen voor een panel van verschillende individuen en dan de sequenties van deze panelleden te vergelijken kun je de kleine verschillen waarnemen.

In Hoofdstuk 2 is deze methode gebruikt voor het vergelijken van acht dieren van verschillende varkenslijnen, te weten een Chinese Meishan, een Belgische Pietrain en een Zweeds wild zwijn en een vijftal Nederlandse witte lijnen.

Zo vergeleken we voor deze acht dieren de sequenties van 67 PCR producten, die samen 32 duizend basen lang waren. Dit komt ongeveer overeen met 0,001 % van het totale genoom. Er werden meer dan 300 SNPs gevonden, een gemiddelde van 1 base verschil per 108 basen.

Van deze SNPs had 63% verschillende allelen binnen de vijf witte lijnen. Deze verschillen zijn dus mogelijk bruikbaar als merker binnen deze commerciële lijnen. Het vergelijken van deze witte dieren met de minder verwante andere dieren leverde hogere percentages op: met Meishan was 89% verschillend, met Pietrain 69% en met wild zwijn 70%. Deze SNPs zijn mogelijk goed bruikbaar in experimentele kruisingen tussen verschillende lijnen, zoals onder andere de Wageningese Meishan X Wit populatie.

Het analyseren en vergelijken van de sequenties bestaat uit diverse opeenvolgende subprocedures. Hoewel een aantal stappen gebruik maakt van specifieke computerprogramma's (waaronder Phred, Phrap en primer3), is het gehele traject een tijdrovende klus. Daarom werd gezocht naar een betere automatisering van dit proces, waarbij de onderzoeker nog wel zeer flexibel de analyse-pijplijn kon samenstellen en ook desgewenst de gebruikte parameters kan aanpassen. Dit resulteerde in twee Perl objecten, samen POSA geheten, en deze staan beschreven in Hoofdstuk 3.

Met behulp van POSA is het mogelijk om in korte Perl-scriptjes alle in de analyse benodigde stappen te definiëren en de gewenste parameters mee te geven. Kort samengevat specificeert de onderzoeker een lijst van de te verwerken (sequentie)bestanden en stelt een kort Perl-script samen waarin de te verrichten analysestappen staan gespecificeerd. De invoer kan bijvoorbeeld uit ruwe data bestaan terwijl de uitvoer kan bestaan uit een lijst met mogelijke base-verschillen (SNPs) of voorstellen voor PCR primers. Al met al kan zo de analysetijd worden teruggebracht van vele uren naar enkele minuten.

SNPs worden veelvuldig gebruikt als genetische markers. Hiervoor dient in een groep individuen (bijvoorbeeld een familie of een groep varkens) bepaald te worden welke allelen ze hebben op de polymorfe baseplaats. Dit proces wordt genotyperen genoemd. Hoewel het onderscheiden van verschil van een enkele base in de twee maal drie miljard basen van het genoom een nauwlettende procedure is, zijn er hiervoor vele methoden beschikbaar. Veel van deze methodes zijn echter kostbaar in arbeidstijd, benodigde apparatuur of andere benodigdheden. Voor laboratoria die over een apparaat om te sequencen (sequencer) beschikken, is een *single-base-extension* (SBE) assay een goede optie, omdat de eind-producten (de genotypes) goed op de sequencer aangetoond kunnen worden.

De wat oudere gel-gebaseerde sequencers worden door een Apple Macintosh computer aangestuurd, en hiervoor is geen goede software beschikbaar om SBE gegevens te

analyseren. Om dit toch te kunnen doen is SNPtyper ontwikkeld, een spreadsheet die de ruwe SBE data van een gel-gebaseerde sequencer zo herschikt en grafisch weergeeft dat deze wel makkelijk interpreteerbaar zijn. SNPtyper maakt het zo toch mogelijk om op de oudere sequencers SBE assays te draaien. Dit is beschreven in Hoofdstuk 4.

Enige jaren geleden waren er diverse experimentele kruisingen opgezet tussen commerciële witte lijnen en Chinese Meishans (beroemd om hun overmaat aan vet; in Wageningen), tussen witte lijnen en Pietrain (Belgische lijn, beroemd om hun gespierdheid/bevleesdheid; in Luik, België) en tussen witte lijnen en wild zwijn (de niet gedomesticeerde voorouder van het huidige varken; in Uppsala, Zweden). In alledrie de kruisingen bleek via een methode die QTL mapping heet, dat er op de korte arm van chromosoom 2 een gen moest liggen dat invloed heeft op de spier- en vetaanzet. Hoewel het gen zelf nog niet geïdentificeerd was, was wel al duidelijk dat alleen de kopie die van de vader afkomstig is tot uitdrukking komt (paternale expressie). De kopie van moeder's kant wordt niet gebruikt (maternal imprinting).

Onlangs is in de Belgische en de Zweedse kruisingen aangetoond dat het betrokken gen insulin-like growth factor 2 (IGF2) heet. Ook werd aangetoond dat het verschil in spier- en spekdikte bepaald wordt door een verschil van een enkele base (te weten IGF2-intron3-G3072A). De base verandering leidde niet, zoals aanvankelijk verwacht, tot een structurele mutatie en een veranderde werking van het eiwit IGF2 zelf, maar bleek een regulatoire mutatie die invloed heeft op wanneer en waar het eiwit zijn werk doet.

Uiteraard was vervolgens de vraag of dezelfde mutatie ook in de Wageningse kruising voorkwam, en zo ja, of deze ook hetzelfde effect op spekdikte heeft. In Hoofdstuk 5 wordt gerapporteerd dat dezelfde mutatie inderdaad in de Wageningse kruising voorkomt en daarin ook een grote invloed heeft op de spekdikte.

Het genoom bestaat uit een aantal chromosomen, die elk tweemaal in ieder lichaamsscel voorkomen. Hoewel deze chromosomen over het algemeen een degelijk en constant geheel vormen, wisselen de twee kopieën van het zelfde chromosoom regelmatig overeenkomstige stukken DNA uit. Dit wordt crossing-over of recombinatie genoemd. Nu is het een gegeven dat hoe verder twee genen of merkers uit elkaar liggen, hoe groter de kans dat er tussen beiden een recombinatie plaatsvindt. Omgekeerd zal een combinatie van twee zeer dicht bij elkaar gelegen merker-varianten (allelen) slechts zeer zelden door recombinatie worden gescheiden en dus (bijna) altijd samen overerven naar de volgende generatie.

Als gevolg hiervan zullen sommige combinaties van allelen vaker voorkomen dan men zou kunnen verwachten op basis van de willekeurige verdeling. Dit verschijnsel noemt men Linkage Disequilibrium (LD). Op basis hiervan kan men voorspellen welke variant van merker 2 een individu heeft als een bepaalde variant bij nabijgelegen merker 1 is

aangetroffen. Merker 1 heeft dusvoorspellende waarde voor merker 2. In Hoofdstuk 6 wordt ingegaan op LD in het varkensgenoom.

De cruciale vraag hierin is nu tot welke afstand merkers nog dicht genoeg bijeen liggen om nog informatief voor elkaar te zijn. Om dit te bepalen is in Hoofdstuk 6 een groot aantal merkers op verschillende onderlinge afstanden gegenotypeerd. Deze gegevens bevatten de allelen op beide chromosoomkopien, en moet eerst worden gesplitst in de allelen op beide losse chromosomen, deze worden haplotypes genoemd. Vervolgens wordt er voor alle mogelijke paren van merkers berekend hoe sterk ze in LD zijn, uitgedrukt op een schaal 0 tot 1. De gegevens in Hoofdstuk 6 laten zien dat de helft van de paren van SNP merkers die tot 50000 basen uit elkaar liggen nog voldoende, maar niet volledige, voorspellende waarde hebben voor elkaar.

In Hoofdstuk 7 worden de belangrijkste bevindingen van dit onderzoek nog eens aangestipt en bediscussieerd in drie secties. De eerste sectie behandelt de methodes voor het identificeren en genotyperen van SNP en het gebruik van SNPs als merkers. Daarnaast wordt ook aandacht geschonken aan de verschillen tussen SNPs en de veelgebruikte microsatelliet merker. De tweede sectie behandelt de implicaties van SNPs in QTL mapping en in het bijzonder de impact die de SNP in het derde intron van IGF2 heeft op de kenmerken vlees- en vetaanzet. In de derde sectie tenslotte wordt LD besproken in het licht van associatie studies.

Het doel van het in dit proefschrift beschreven onderzoek was de implementatie en toepassing van SNPs als merkers in fokkerij. De nadruk lag hierbij vooral op analyse van vetkenmerken bij varkens, en dan specifiek op de ingeprente QTL regio op chromosoom 2. Diverse aspecten van SNPs als merkers zijn hierbij aan bod gekomen, waaronder methodes en gereedschappen om SNPs te identificeren en te genotyperen. Daarnaast werd de QTL regio op SSC2 opnieuw geanalyseerd na de identificatie van een SNP in het IGF2 gen die een grote invloed had op vlees- en vetaanzet. Hierin werd aangetoond dat, ook in de Wageningse kruising, de QTL regio werd teruggebracht tot een enkele base. Voor verder onderzoek in andere genomische regio's is het van belang te kunnen inschatten hoeveel genetische merkers (zoals SNPs) er nodig zullen zijn in associatie studies. Om hiervan een goede schatting te kunnen maken, werd het patroon van linkage disequilibrium in het varkensgenoom bestudeerd. Het bleek dat de meerderheid van de SNP-paren die tot 50 kb uit elkaar liggen nog voldoende mate van LD vertonen om in associatie studies gebruikt te kunnen worden. Verder onderzoek hiernaar blijft evenwel noodzakelijk, omdat dit slechts een gemiddelde betreft, en de patronen lokaal erg uiteen kunnen lopen.



## **List of publications**



**BJ Jungerius**, AP Rattink, R Crooijmans, J van der Poel, BA van Oost, MFW te Pas, MAM Groenen (2003) Development of a SNP map of porcine chromosome 2. *Animal Genetics*, 34: 429-437.

**BJ Jungerius**, A Veenendaal, BA van Oost, MFW te Pas, MAM Groenen. (2003) Typing single nucleotide polymorphisms using a gel based sequencer: an improved efficiency protocol and a new data analysis. *Molecular Biotechnology*, 25(3): 283-287.

**BJ Jungerius**, A-S vanLaere, MFW te Pas, BA van Oost, L Andersson, MAM Groenen (2004) The IGF2-intron3-G3073A substitution completely explains a major imprinted QTL effect on backfat thickness in a Meishan X European white pig intercross. Accepted for publication in *Genetical Research*.

**BJ Jungerius**, J Gu, BA van Oost, MFW te Pas, MAM Groenen (2004) An estimation of the extent of Linkage Disequilibrium in the porcine genome. Submitted.

JA Aerts, **BJ Jungerius**, MAM Groenen (2004) Read.pm and Contig.pm - Perl objects to facilitate handling of sequencing reads and to automate SNP discovery. *BMC Genomics* 5: 60..

AP Rattink, **BJ Jungerius**, M Faivre, P Chardon, B Harlizius and MAM Groenen. (2001) Improving the comparative map of SSC2p-q13 by sample sequencing of BAC clones. *Animal Genetics*, 32(5):274-80.

AP Rattink, M Faivre, **BJ Jungerius**, MAM Groenen, B Harlizius (2001) A high-resolution comparative RH map of porcine chromosome (SSC) 2. *Mammalian Genome* 12(5):366-70.



## **Nawoord**



Zo.

En dit is het dan geworden.

Een verslag over het onderzoek waar ik dik vier jaar mee bezig ben geweest.

Gelukkig hoefde ik niet alles alleen te doen. Er zijn veel personen zonder wie dit boekje er niet zo zijn gekomen, of zonder wie het er in ieder geval anders zou hebben uitgezien.

Voor al deze personen: Bedankt!

Toch wil ik een aantal personen met name noemen.

In de eerste plaats de heren (co-)promoteren: Martien, Bernard en Marinus. Onze tweemaandelijksse bijeenkomsten heb ik als nuttig ervaren, juist de verschillen in jullie visies leidde tot discussie. En zelfs met drie projectleiders was de inhoudelijke vrijheid bij het invullen van het project erg prettig.

Barbara, in het eerste jaar mijn kamergenoot. Doordat je de dagelijkse begeleiding deed, waren je interesse, hulp en ervaring een extra stimulans. In ieder geval heeft het me een eind op weg geholpen.

Two MSc students participated in this research for their theses. First, Cleopatra explored the feasibility for investigating porcine LD. Then Jingjing took over and performed the main part of the work for the LD study. Thank you both for your efforts, as it is a valuable chapter in this thesis!

Tineke, vooral bij de eerste hoofdstukken was je hulp en meedenken onmisbaar en Jan (A.), vooral bij de laatste hoofdstukken was je hulp en meedenken onmisbaar. Beiden ook bedankt voor de (on)nuttige gesprekken met (Vlaams) bier en alvast bedankt dat jullie bij de verdediging naast me willen zitten...

Maria bedankt voor de dropjes en de hulp bij QTL analyses (in volgorde van belangrijkheid). Tarik, as roommates we often discussed science, regulations, bureaucracy and the weather. Thanks for pleasant company.

Verder wil ik ook mijn familie en vrienden bedanken voor de interesse die ze altijd getoond hebben. En tot slot een woord van dank voor Annemieke voor alle hulp, support en afleiding, zowel binnen als buiten de vakgroep.



Bart





# Curriculum Vitae



Bart Johan Jungerius werd op 1 februari 1973 geboren te Maastricht en groeide op in Margraten. In 1991 behaalde hij het VWO-diploma aan het Sint-Maartenscollege te Maastricht. In september van hetzelfde jaar begon hij aan de toenmalige Landbouwuniversiteit Wageningen (LUW) met een studie Moleculaire Wetenschappen. Tijdens zijn eerste afstudeervak cloneerde hij bij de sectie Moleculaire Genetica van de vakgroep Erfelijkheidslere van het Nrf1 gen van de schimmel *Cladosporium fulvum*. Het tweede afstudeervak werd uitgevoerd bij de vakgroep Biochemie en was gewijd aan het enzym Pyruvate:ferredoxine oxidoreductase van het hyperthermofiele organisme *Pyrococcus furiosus*. Tot slot volgde een afstudeervak bij de divisie Immunologische en Infectieziekten van TNO Preventie en Gezondheid, waar hij werkte aan de heterologe expressie van rotavirale wandeiwitten in *E.coli*. In 1998 runde hij zijn studie af.


Vanaf september 1999 werkte hij als AIO aan het in dit proefschrift beschreven onderzoek. Dit onderzoek was een gezamenlijk project van de faculteit Diergeneeskunde van Universiteit Utrecht, de leerstoelgroep Fokkerij en Genetica van Wageningen Universiteit (momenteel onderdeel van Animal Sciences Group (ASG) van Wageningen Universiteit en Research centre (WUR)), en het toenmalige ID-DLO (later ID-Lelystad, momenteel onderdeel van ASG van WUR).

Per 1 augustus 2004 is hij werkzaam als postdoc bij de Sectie Complexe Genetica van de Divisie Biomedische Genetica van het Universitair Medisch Centrum Utrecht.



## **Training and Supervision Plan**



Training and Supervision Plan		Graduate School WIAS	
Name PhD student	Bart Jungerius		
Project title	Development of high-throughput multiplex analysis of genes affecting fatness traits in the pig.		
Group	Animal Breeding & Genetics		
Supervisor(s)	Martien Groenen (ABG), Bernard van Oost (UU) and Marinus te PAS (ID)		
Project term	01-09-1999 - 31-10-2003		
Education and Training (minimum 21 cp, maximum 42 cp)		year	cp
<b>The Basic Package</b> (minimum 2 cp)			
WIAS Common Course		1999	2.0
Course on philosophy of science and ethics		2000	1.0
Subtotal			3.0
<b>Scientific Exposure</b> (minimum 5 cp)			
<i>International conferences</i> (minimum 2 cp)			
International Society for Animal Genetics, Göttingen, D.		2002	1.0
Plant and Animal Genome Conference, San Diego, CA, USA. (2002, 2003)			2.0
NWO-MW Genetics Retraite Rolduc, NL. (1999, 2000, 2001, 2002)			1.6
<i>Seminars and workshops</i>			
WIAS Science Days (2000, 2001, 2003)			0.6
array-NL platform meetings (2000, 2001)			0.6
UU-ABC-SNP meeting, Utrecht		2001	0.2
Wageningse Kennisdagen		2000	0.2
<i>Presentations</i>			
International Society for Animal Genetics, Göttingen, D.		2002	0.5
Plant and Animal Genome Conference, San Diego, CA, USA. (2002, 2003)			1.0
NWO-MW Genetics Retraite Rolduc, NL.		2001	0.5
UU-ABC-SNP meeting, Utrecht		2001	0.5
Subtotal			8.7
<b>In-Depth Studies</b> (minimum 4 cp)			
HGMP-RC course; Staden Package, Cambridge, UK.		2000	0.5
Acquiring and Analyzing Genomic Sequence Data, Cold Spring Harbour, NY, USA.		2002	4.0
ToPigs, discussions with all WIAS 'Pig'Phds. (1999-2001)			1.0
Winterschool Bioinformatics, WUR.		2000	1.0
Subtotal			6.5
<b>Professional Skills Support Courses</b> (minimum 2 cp)			
WIAS Course Techniques for Scientific Writing		2002	0.8
Use of Laboratory Animals		2000	3.0
Laboratory Use of Isotopes		2001	1.0
Subtotal			4.8
<b>Didactic Skills</b>			
Werkcollege Genetica, Utrecht		2001	1.0
Werkcollege Genomics, WU		2003	2.0
Werkcollege Toegepaste Dierbiologie, WU		2002	0.4
Practicum Inleiding Fokkerij (2000, 2001)			0.4
Supervising MSC theses (2), WU			2.0
Subtotal			5.8
<b>Management Skills</b>			
Organisatie Cursus Proefdierkunde, WU.		2001	4.0
AIO-vertegenwoordiger stafoverleg leerstoelgroep (2001-2003)			0.5
Subtotal			4.5
<b>Total</b> (minimum 21 cp, maximum 42 cp)			33.3

