Mammalian Genome 13, 569–577 (2002).
DOI: 10.1007/s00335-002-2184-5

Mammalian Genome
Incorporating Mouse Genome

A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking *KIT*

Elisabetta Giuffra,[1,*] Anna Törnsten,[1] Stefan Marklund,[1] Erik Bongcam-Rudloff,[1] Patrick Chardon,[2] James M.H. Kijas,[1,] Susan I. Anderson,[3] Alan L. Archibald,[3] Leif Andersson[1]**

[1]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 597, S-751 24 Uppsala, Sweden
[2]Laboratoire de Radiobiologie et d'Etude du Génome, CEA INRA, F-78352 Jouy-en-Josas Cedex, France
[3]Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, Scotland, UK

Received: 20 March 2002 / Accepted: 9 July 2002

Abstract. The *Dominant White* (*I*/*KIT*) locus is one of the major coat color loci in the pig. Previous studies showed that the *Dominant White* (*I*) and *Patch* (I^P) alleles are both associated with a duplication including the entire *KIT* coding sequence. We have now constructed a BAC contig spanning the three closely linked tyrosine kinase receptor genes *PDGFRA–KIT–KDR*. The size of the duplication was estimated at about 450 kb and includes *KIT*, but not *PDGFRA* and *KDR*. Sequence analysis revealed that the duplication arose by unequal homologous recombination between two LINE elements flanking *KIT*. The same unique duplication breakpoint was identified in animals carrying the *I* and I^P alleles across breeds, implying that *Dominant White* and *Patch* alleles are descendants of a single duplication event. An unexpected finding was that Piétrain pigs carry the *KIT* duplication, since this breed was previously assumed to be wild type at this locus. Comparative sequence analysis indicated that the distinct phenotypic effect of the duplication occurs because the duplicated copy lacks some regulatory elements located more than 150 kb upstream of *KIT* exon 1 and necessary for normal *KIT* expression.

Introduction

Dominant White (*I*) is one of the major coat color loci in the pig and is equivalent to the *KIT* locus. *KIT* encodes the mast/stem cell growth factor receptor. KIT and its ligand (MGF) have a crucial role for the survival and migration of neural-crest-derived melanocyte precursors, and for the development of hematopoietic cells, primordial germ cells, and interstitial cells in the small intestine. Structural and regulatory *KIT* mutations cause dominant white spotting in a number of mammalian species including mouse (MGI:96677, http://www.informatics.jax.org), human (OMIM *164920, http://www.ncbi.nlm.nih.gov), pig (Giuffra et al. 1999; Johansson Moller et al. 1996; Marklund et al. 1998), and most likely also horse (Marklund et al. 1999) and cattle (Reinsch et al. 1999). In the mouse, loss-of-function mutations are associated with limited white spotting (often a belly spot and spot in the forehead) in heterozygotes, but they are often lethal or sublethal in the homozygous condition owing to pleiotropic effects on hematopoiesis. Many regulatory mutations are dominant and have more drastic effects on pigmentation in heterozygotes, causing large areas of white coat as observed in *Sash* (W^{sh}), *Banded* (W^{bd}), and *Patch* (*Ph*). Regulatory mutations primarily affecting melanocyte development may be fully viable in the homozygous condition. Several of the regulatory mouse mutations are associated with chromosomal rearrangements that affect presumed regulatory elements located 20–200 kb upstream of *Kit* (Berrozpe et al. 1999; Hough et al. 1998).

Four different *Dominant White*/*KIT* alleles with distinct phenotypic effects have been described so far (Giuffra et al. 1999; Johansson Moller et al. 1996; Marklund et al. 1998); the wild type (*i*), present in the Wild Boar and in colored breeds; *Patch* (I^P), causing patches of white color, found in Landrace and Large White pigs; *Dominant White* (*I*) causing a fully dominant white color in Landrace and Large White pigs; and *Belt* (I^{Be}), causing a white belt across the shoulders and front legs in Hampshire pigs and most likely in other breeds with the Belt phenotype. *I* and I^P are associated with a duplication of the entire *KIT* coding sequence (Johansson Moller et al. 1996). We assume that the duplication acts as a regulatory mutation and that the phenotypic effect is due to overexpression or ectopic expression of KIT. The *Dominant White* (*I*) allele carries, in addition, a splice mutation at the first nucleotide of intron 17 in one of the two *KIT* copies. The splice mutation leads to skipping of exon 17 and the expression of a truncated form of KIT that is expected to lack tyrosine kinase signaling (Marklund et al. 1998). Our previous RT-PCR analysis and immunocytochemistry showed that both *KIT* copies are expressed in embryonic and adult tissues. Thus, the fully dominant white phenotype is due to the combined effect of a regulatory mutation (the duplication) and the structural splice mutation. The *Belt* allele contains a single *KIT* copy and is most likely caused by a regulatory mutation (Giuffra et al. 1999).

The aim of this study was to construct a BAC contig of the *Dominant White*/*KIT* locus and to clone and characterize the duplication breakpoint. We report here that the duplication is about 450 kb in size and occurred by unequal homologous recombination between two LINE elements flanking *KIT*.

Materials and methods

Pulsed field gel electrophoresis (PFGE) and Southern blot analysis. DNA plugs were prepared from fresh or frozen blood of Duroc (*i*/*i*),

* *Present address:* Centro Ricerche Studi Agroalimentari FPTP-CERSA, LITA, Via Fratelli Cervi 93, 20090 Segrate, Italy.
** *Present address:* Baker Institute of Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853, USA.
Correspondence to: L. Andersson; Email: Leif.Andersson@bmc.uu.se, Phone +46-18 4714904, Fax +46-18 4714833,
The sequence data described in this paper have been submitted to GenBank with accession numbers AF378821–AF378823, BH011551–BH011594.

Hampshire (I^{Be}/I^{Be}), and Large White (I/I) pigs. White cells were prepared by isotonic lysis, washed two to three times in PBS, resuspended at a concentration of 25×10^6 cells/ml, and mixed with an equal volume of 1.5% low-melting agarose in PBS cooled to 50°C. Aliquots of the agarose-cell suspension were placed in plug molds (Bio-Rad Laboratories, Hercules, Calif.) and allowed to solidify at 4°C. Plugs were digested for 1–2 days at 50°C with constant shaking in 0.5 M EDTA, pH 8.0, 1% Sarkosyl, 0.5 mg/ml Proteinase K. After equilibration in TE, plugs were incubated in TE containing 1 mM phenylmethylsulfonyl fluoride to inactivate residual Proteinase K activity. After extensive washing, the plugs were stored at 4°C in 0.5 M EDTA or used directly for restriction digestion. Each plug was divided into two parts of approximately 35 µl and equilibrated for about 3–4 h on ice in the restriction buffer provided by the manufacturer (New England Biolabs Inc., Beverly, Mass.). The buffer was replaced by fresh buffer containing about 50 U of enzyme, and the enzyme was allowed to diffuse into the plug for 16 h at 16°C. After incubation at 37°C for 16 h, about 20 U of enzyme was added, allowed to diffuse into the plug for 4–5 h at 16°C, and incubated at 37°C for an additional 5–6 h.

PFGE of the digested plugs was performed in a CHEF Mapper XA apparatus (Bio-Rad Laboratories) at 14°C in a 1% agarose gel in 0.5 × TBE. Electrophoresis conditions were set by the Auto Algorithm Mode to obtain the optimal resolution for the expected fragment sizes, typically between 50 and 800 kb (pulse times of 6 s to 1.3 min, in an electric field of 6 V/cm for 27 h). Yeast chromosomes and Lambda ladder (Bio-Rad Laboratories) were used as size markers. DNA separated by PFGE was transferred to Hybond N+ membranes (Amersham Pharmacia Biotech, Uppsala, Sweden).

Hybridizations were performed in ExpressHyb hybridization solution (Clontech, Laboratories Inc., Palo Alto, Calif.). The DNA probes used were: a 2.4-kb *BamH*I/*Sal*I fragment of pig *KIT* cDNA (Marklund et al. 1998); a 3.4-kb *BamH*I fragment of human *PDGFRA* cDNA (Claesson-Welsh et al. 1989); a 4.5-kb *Xho*I/*Xba*I fragment of human *KDR* cDNA (GenBank AF063658); a 229-bp *KIT* intron 18 PCR fragment (Johansson Moller et al. 1996); and four STS probes from the BAC contig: STS 1000D25′, STS 211E125′, STS 645D53′, and STS 953F113′. Probes were labeled with ^{32}P-dCTP by using the Megaprime DNA Labelling System (Amersham Pharmacia Biotech). Hybridized blots were exposed in a PhosphorImager (Molecular Dynamics, Sunnyvale, Calif.) for at least 16 h.

Southern blot analysis of *Hind*III-digested BAC DNA and genomic DNA was carried out as previously described (Johansson Moller et al. 1996). An 897-bp PCR product from the 3.9-kb subclone pUC953H4 containing the 3′-5′ duplication breakpoint was used as probe.

BAC contig. The porcine BAC library (Rogel-Gaillard et al. 1999) was screened by PCR, and the BAC end sequences were determined by direct sequencing of both the 5′ and 3′ ends of selected BACs as described (Jeon et al. 2001). The BAC end sequences defined new Sequence Tagged Sites (STSs) that were used to screen the library again and expand the contig. All gene-specific and STS primers used in this study are provided in Table 4. The order and overlap of the BAC clones were determined by screening the STSs against all clones in the contig. The overlap and physical distances between BAC clones were also estimated by restriction mapping. The rare cutting enzymes *Sma*I and *Not*I (Amersham Pharmacia Biotech) were used for complete and partial digestions. FISH analysis of BAC clones was carried out as previously described (Chowdhary et al. 1995).

The obtained BAC sequences were masked for interspersed repeats and low complexity DNA sequences using RepeatMasker (A.F.A. Smit and P. Green, unpublished data; http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). The mammalian library of repeats provided with the program was updated with a consensus pig SINE sequence and other pig-specific repeats. Masked sequences were subjected to BLAST (Altschul et al. 1990) searches against DNA databases (nr, month, and dbest) at NCBI by using the advanced BLAST version 2.0 network service (http://www.ncbi.nlm.nih.gov/). Subsequently, masked and unmasked sequences were used to screen the non-redundant tiling path of the human draft genome sequence at http://www.enseml.org.

PCR amplification of the duplication breakpoints. Fragments spanning the breakpoints were amplified from pig genomic DNA samples representing the Wild Boar, Large White, Landrace, Piétrain, Berkshire, Duroc, Hampshire, Linderöd, and Meishan breeds. The primers 832F2 (5′-CCA CAA TAT ACC TAC CAG AAT TAC) and 953R2 (5′-AAC CTG TGG ATC AAA TCT GGT C) were used to amplify 968 bp spanning the 5' breakpoint present in BAC832E11; 953F1 (5′-GTT CAA TCC AGC AAT CAC AAC C) and 953R2 were used to amplify 864 bp spanning the 3′-5′ breakpoint present in BAC953F11; and 953F1 and 1041R1 (5′-TTT TAA TCC TCT TAA GGA CCA AC) were used to amplify 1022 bp spanning the 3' breakpoint present in BAC 1041B3. The PCR reactions were performed in 10-µl reactions including 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, 2.5 pmol of each primer, 5% DMSO, 25 ng genomic DNA, 1× PCR GOLD buffer, and 0.75 U Ampli*Taq* GOLD polymerase (PE Applied Biosystems, Foster City, Calif.). Thermocycling was carried out with a PTC 200 instrument (MJ Research, Watertown, Mass.). The temperature conditions in the first cycle were 94°C for 10 min, 55°C for 30 s, and 72°C for 90 s, whereas the remaining cycles were performed at 94°C for 30 s, 52°C for 30 s, and 72°C for 90 s. The PCR products were directly sequenced as described above.

The primers 953F9 (5′-TAA GTG AAA GAA GTC AAT CTG AG) and 953R3 (5′-GGC AGT CAT GTA ACT ATC ACC) were used to generate a 152-bp product spanning the 3'-5′ breakpoint as a diagnostic test for the duplication. The product was separated by standard agarose gel electrophoresis.

Results

Pulsed field gel electrophoresis (PFGE). DNA samples from Duroc, Hampshire, and Large White pigs representing the three *Dominant White*/*KIT* genotypes i/i, I^{Be}/I^{Be}, and I/I, respectively, were used. DNA plugs were digested with three rare-cutting enzymes *Nar*I, *BssH*II, and *Pme*I. The digested DNAs were separated by PFGE and blotted to hybridization membranes. The membranes were first hybridized with *KIT* probes and cDNA probes for *PDGFRA* and *KDR*, which are flanking *KIT* in the human genome (Spritz et al. 1994). The results showed, as expected, that *KIT* hybridized to the duplicated region since two to three fragments were obtained for both *Nar*I and *BssH*II in Large White pigs, but only a single fragment in non-white pigs (Fig. 1A; Table 1). The *PDGFRA* probe hybridized to a single fragment with all enzymes and in all genotypes, indicating that the duplication does not involve this gene, in agreement with our previous FISH analysis (Johansson Moller et al. 1996). Similarly, the *KDR* probe hybridized to a single fragment in Large White pigs with all three enzymes.

The results indicated that one duplication breakpoint is located between *PDGFRA* and *KIT*, while another is located between *KIT* and *KDR*. Two STS fragments 645D53′ and 211E125′ from the region between *KIT* and *PDGFRA* were hybridized to the PFGE blots. The results showed that the breakpoint is located between these two STSs, since only the latter hybridized to duplicated fragments (Fig. 1B; Table 1). The observation of three restriction fragments for some enzyme/probe combinations may be explained by restriction site polymorphisms, incomplete digestions, or the presence of haplotypes with three *KIT* copies as recently documented to occur quite frequently in white breeds (Pielberg et al. 2002). It is also worth noticing that the *Belt* allele (I^{Be}) present in Hampshire pigs was associated with a difference in restriction fragment sizes compared with Duroc (i/i) with all three restriction enzymes when the *KIT* intron 18 probe was used (Table 1). The observation may be relevant for the identification of the causative mutation for *Belt* as several *Kit* alleles in mice are due to chromosomal rearrangements (Berrozpe et al. 1999; Hough et al. 1998).

BAC contig. The BAC library was constructed from a Large White pig assumed to be homozygous I/I. We have arbitrarily designated the normal gene copy *KIT1* and the copy contain-

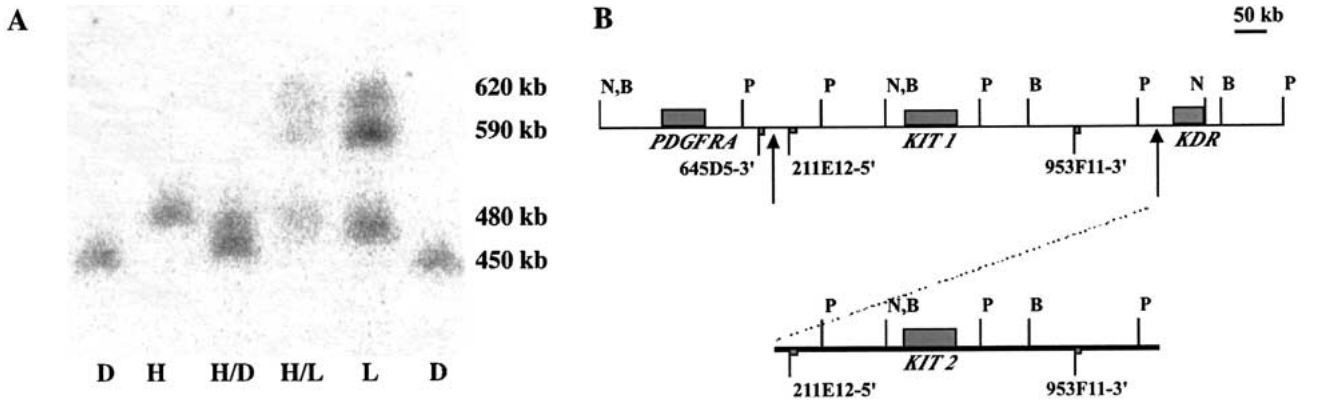


Fig. 1. Results of PFGE analysis of the *PDGFRA–KIT–KDR* region in pigs. (**A**) Southern blot analysis of *Nar*I-digested genomic DNA hybridized to a *KIT* intron 18 probe. Samples from Duroc (D) *KIT-i*/*i*, Hampshire (H) *KIT-*I^{Be}/I^{Be}, Large White (LW) *KIT-I*/*I*, a Duroc/Hampshire crossbred animal, and a Hampshire/Large White crossbred animal were used. The estimated sizes of fragments are given to the right in kilobase pairs. (**B**) Schematic figure summarizing the interpretation of the PFGE data with the non-duplicated chromosome above and the duplication below. The locations of restriction sites are indicated by vertical lines; B = *BssH*II, N = *Nar*I, and P = *Pme*I. The approximate locations of the duplication breakpoints are indicated by arrows, and the locations of the gene and STS probes used for hybridization are indicated by boxes.

Table 1. Restriction fragment sizes observed in PFGE analysis of the *PDGFRA–KIT–KDR* region on pig Chromosome 8.

		Probe					
Restriction enzyme	Breed	*PDGFRA* CDNA	*STS* 645D53′	STS 211E125′	*KIT* intron 18 (and cDNA)	STS 953F113′	*KDR* cDNA
*Nar*I	D	450	450	450	450	-	450
	H	450	450	450	480	-	480
	LW	450	450	450,590,620	480,590,620	-	480
*BssH*II	D	450	450	450	200	320	320
	H	450	450	450	225	320	320
	LW	450	450	450,400	210,215,225	320,400	320
*Pme*I	D	-	130	130	410	410	225
	H	-	130	130	240	240	225
	LW	-	130	130,100	240	240	225

D = Duroc; H = Hampshire; LW = Large White; "-" = not tested.

ing the splice mutation *KIT2*. The construction of the BAC contig was initiated by screening the library with primers amplifying *KIT* exons. We were able to assign these BAC clones as *KIT1* or *KIT2*, using the diagnostic test for the splice mutation (Fig. 2), but we do not know in which order the two copies occur in relation to *PDGFRA* and *KDR*. The contig was expanded on both sides by chromosome walking with STSs developed by BAC end sequencing. Clone 211E12 contains *KIT* exon 1 and parts of the upstream region, while clone 549C3 represents the *KIT* downstream region. We were unable to assign these clones to the *KIT1* or *KIT2* region owing to the lack of diagnostic polymorphisms.

The closely linked *PDGFRA* gene was chosen as a second starting point for building the BAC contig. PCR primers were designed with *PDGFRA* sequences conserved between human and mouse. Two positive clones were identified (Fig. 2). The BAC ends were sequenced and used to develop new STS primers that in turn were used to expand the contig on both sides of *PDGFRA*. Six additional clones were identified and subjected to BAC end sequencing. BLAST searches using the STS 642D43′ revealed a highly significant similarity to the human *PDGFRA* promoter region. The result provided an orientation of the subcontig, and the chromosome walking was continued from the 3′-end of *PDGFRA* with the assumption that *PDGFRA* and *KIT* are oriented head-to-tail in pigs as in humans (Spritz et al. 1994). BAC screening with STS 645D53′ identified two new clones; one of these, 832E11, overlapped with clone 211E12, showing that the *PDGFRA-KIT* contig was closed (Fig. 2). The distance between the two genes was estimated at 350 kb, very similar to the corresponding estimate for the mouse (Hough et al. 1998).

The PFGE data indicated that one duplication breakpoint should be located between STS 645D53′ and 211E125′. These two STSs are about 100 kb apart and both present in BAC 832E11, which should thus contain the 5′ breakpoint. The BAC library was, therefore, screened with STS 211E125′ on the assumption that this STS should be able to identify BACs from the 5′ breakpoint as well as from the 3′-5′ breakpoint (see Fig. 2). The characteristic feature of the latter types of clones would be that they should contain one end not belonging to the *PDGFRA-KIT* interval. Three new clones were isolated by using STS 211E125′, and new STSs were generated by BAC-end sequencing. PCR screening of the BACs from the contig revealed that both 763F13′ and 953F113′ were only positive with themselves, indicating that the corresponding clones potentially represented the 3′-5′ breakpoint. FISH analysis of 763F1 showed that this clone was chimeric with one end from the *KIT* region Chromosome (Chr) 8 and the other end originating from Chr 5. However, FISH analysis of clone 953F11 only resulted in a signal from the *KIT* region on SSC8q12. STS 953F113′ was then used to isolate four new BACs (391B8, 460F10, 568E1, 1041B3; Fig. 2). BAC-end sequencing and BLAST searches against GenBank revealed a highly significant hit between the 3′-end of clone 1041B3 and the *KDR* coding sequence in different species; the highest score was obtained against human *KDR*, AF063658 (94% sequence identity over 93 nucleotides, P = $1e^{-32}$). This, together with the PFGE data showing that *KDR* is not duplicated, provided evidence that

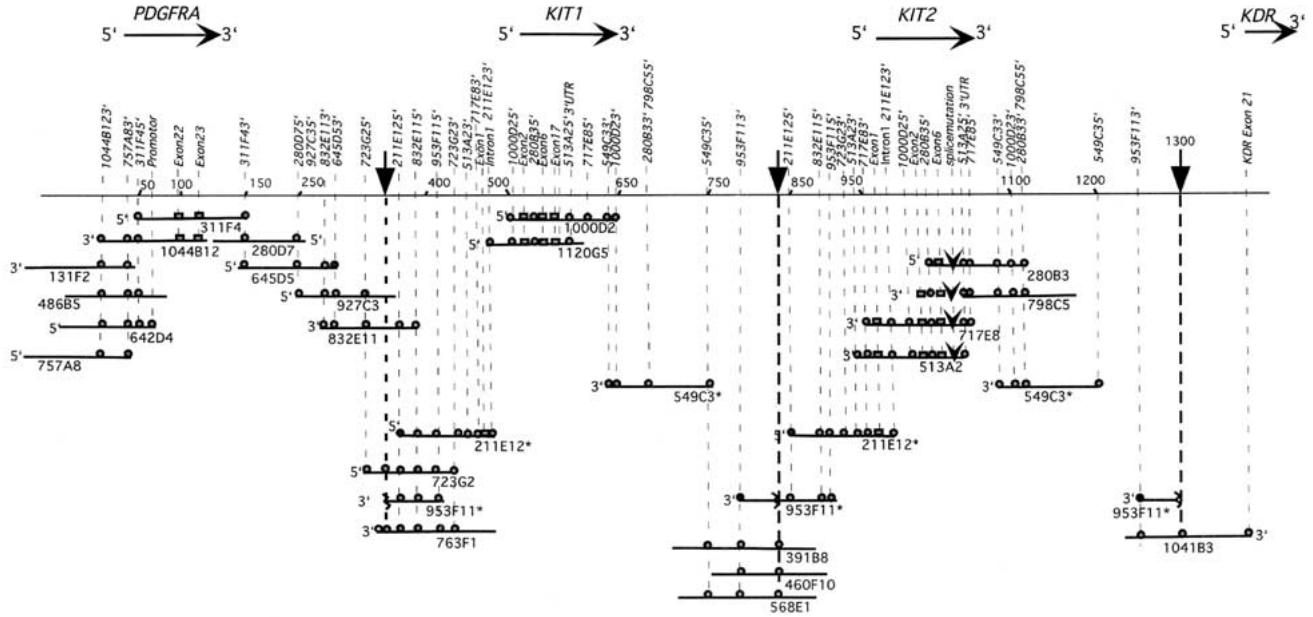


Fig. 2. Map of the BAC contig of the *PDGFRA–KIT–KDR* region on pig Chr 8. The orientation of the genes has been determined for *PDGFRA*, *KIT1*, and *KIT2*, while the orientation of *KDR* is given according to the one established in human (Spritz et al. 1994). The presence of the splice mutation in *KIT2* is marked with an arrowhead. The duplication breakpoints are indicated by vertical arrows. Exons (boxes) and STSs (circles) are indicated. The BAC clones marked by asterisks are present twice because it remains to be established whether they belong to *KIT1* or *KIT2*. The chimeric 3′ end of BAC 763F1 (see text) is marked by an open circle.

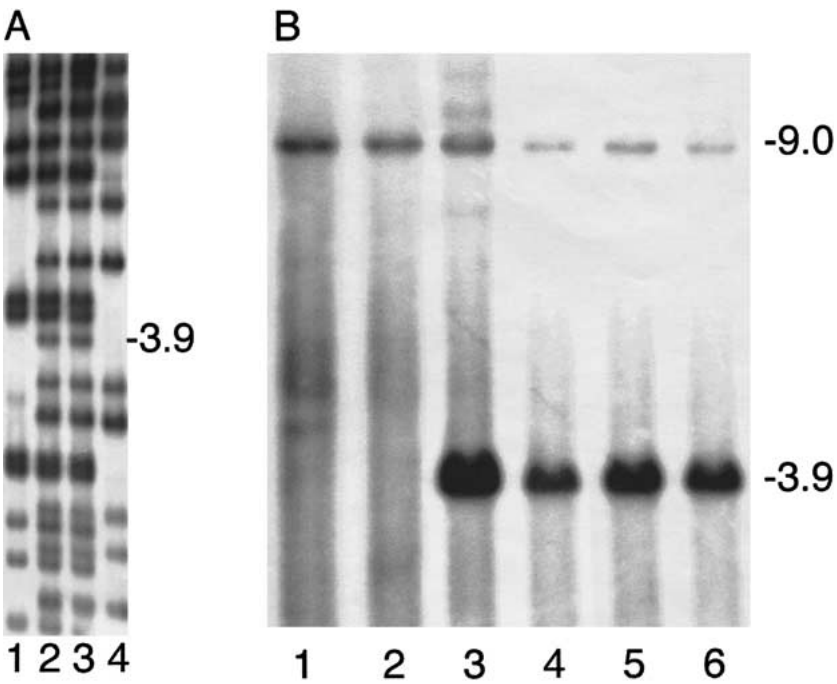
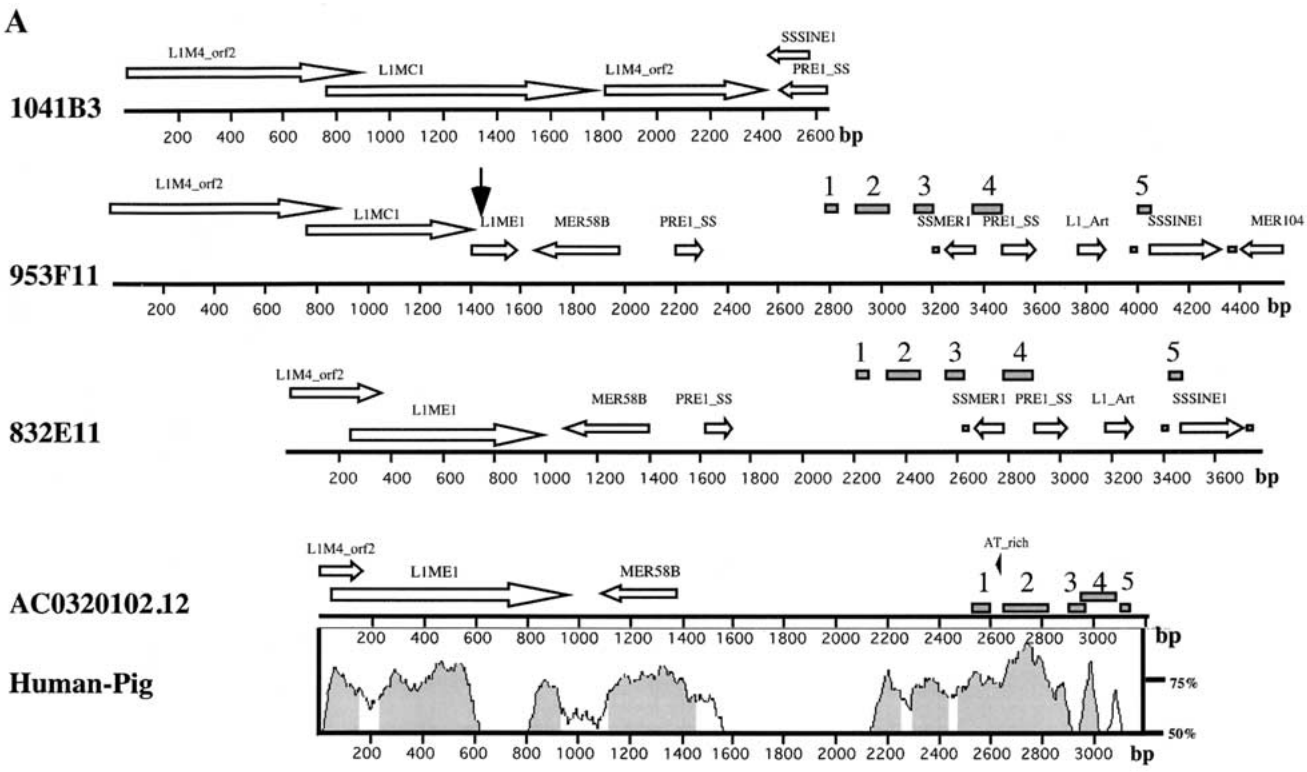


Fig. 3. Southern blot analysis of *Hind*III-digested BAC clones from the *KIT* region and *Hind*III-digested pig genomic DNA. (**A**) BAC clones 832E11 (lane 1), 953F11 (lane 2 and 3), and 1041B3 (lane 4) hybridized with BAC DNA from clone 953F11. The unique 3.9-kb *Hind*III fragment present at the duplication breakpoint is marked. (**B**) Genomic Southern blot of founder animals from a Wild Boar/Large White intercross hybridized with a PCR fragment from the duplicated region. 1: W1, i/i; W2, i/i; W5, I/I; W6, I/I^{Be}; W7, I/I; W8, I/I^{P}. The unique 3.9-kb *Hind*III fragment from the duplication breakpoint is indicated. The 9.0-kb fragment was monomorphic and originates from the region represented in BAC 832E11.

clone 953F11 contains the 3′-5′ duplication breakpoint and that 1041B3 contains the 3′ breakpoint (Fig. 2). PCR screening using the STSs isolated from BAC 549C3 from the 3′-region of *KIT* revealed that the 5′ STS was positive, with 391B8 and 568E1 showing that the *KIT-KDR* contig had been closed. This allowed us to estimate the distance between *KIT* and *KDR* to 250 kb and the size of the entire duplication to about 450 kb.

Characterization of the duplication breakpoint. A Southern blot of *Hind*III-digested DNA from the three BAC clones 832E11, 953F11, and 1041B3 containing the three duplication breakpoints was hybridized with clone 953F11; *Hind*III was chosen, since partial *Hind*III digestions were used for the library construction. BAC 953F11 contained a unique 3.9-kb *Hind*III fragment, whereas all other fragments were also present in 832E11 or 1041B3, or in both, as regards fragments representing the BAC vector (Fig. 3A). The results confirmed our interpretation that BAC953F11 represents the 3′-5′ breakpoint. The duplication appears to be a recent event, since the data show that no *Hind*III restriction site has been gained or lost and no large insertions/deletions have occurred since the duplication event. BAC 953F11 contains a small region



B

```
                10        20        30        40        50        60
953F11    ATCTGAGAAGGCTACATACTGTATGATTCCAAGGGTCATGGCTTGAAAAAGAGACTGACC
1041B3    ------------------------------C-A-ATGACA--CTGG--AT-G-AA-A-
832E11    CAGAT-AC-TAAAG--G-TCAGTG-T-G----------------------------
```

Fig. 4. Sequence analysis of the unique 3′-5′ duplication breakpoint in the porcine *KIT* gene isolated from BAC953F11. The sequence is compared with the corresponding regions in BACs 1041B3 and 832E11. (**A**) Annotation of the sequences. A vertical black arrow indicates the location of the duplication breakpoint. Filled boxes (1–5) indicate sequences with highly significant similarity to the corresponding region in the human genome (sequence contig AC0320010212). The locations of repetitive sequences are indicated by open arrows. LINE repeats: L1M4_orf2, L1MC1, L1ME1, L1ME2 and L1_art (Smit et al. 1995); MER repeats: MER58B (Kaplan et al. 1991), MER104 and SSMER1 (porcine MER1); porcine SINE repeats: PRE1_SS and SSSINE1 (Frengen et al. 1991). The scale provided is in base pairs. A comparative annotation of the corresponding human sequence in contig AC0320010212 is also shown. A sequence identity plot obtained with VISTA (Mayor et al. 2000) at http://www-gsd.lbl.gov for the comparison of the pig and human sequence is shown at the bottom. Percentage identity was calculated with a window length of 100 bp. Conserved sequences are shown relative to their positions in the human sequence, and the percentages of identities (50%–100%) are indicated on the vertical axes. (**B**) Sequence alignment of the 60 bp around the duplication breakpoints in the three BAC clones. A dash indicates identity to the master sequence.

upstream of *KIT* not present in BAC832E11, implying that the 3.9-kb *Hind*III fragment potentially represented this region. However, the subsequent sequence analysis, as well as Southern blot analysis of genomic DNA, convincingly demonstrated that it constitutes the duplication breakpoint (see below).

The 3.9-kb *Hind*III fragment from 953F11 was subcloned into pUC18 to generate clone pUC953H4. The fragment was sequenced with vector primers and primer walking. The corresponding sequences from BAC 832E11 and 1041B3 were generated by PCR amplification and direct sequencing. The sequence comparison confirmed that 953F11 is a hybrid between the sequences of 832E11 and 1041B3. Bioinformatic analysis revealed that the 3.9-kb fragment contains several repetitive elements and that the duplication breakpoint is a hybrid LINE element corresponding to partial LINE elements in 832E11 and 1041B3 (Fig. 4). BLAST analysis showed that a region about 1 kb 3′ of the duplication breakpoint and also present in 832E11 showed several highly significant hits to the corresponding region in the human genome (GenBank AC03201212). The human region is located about 150 kb upstream of *KIT* exon 1, in excellent agreement with the location of this region in the pig (Fig. 2). A comparison of unmasked sequences showed that the L1 repeats and the MER58B repeat located at or very close to the duplication breakpoint were conserved between pigs and human as regards both content and orientation of repeats. The output from the VISTA program reveals a striking overall sequence similarity between pig and human (Fig. 4A).

A sequence comparison of 1,195 bp of 953F11 and the corresponding region in 1041B3 revealed no sequence difference, and a comparison of about 2450 bp in 953F11 and the corresponding region in 832E11 revealed two differences, one extra nucleotide at a mononucleotide repeat and a single base

Table 2. Sequence variation at the 5′ and 3′ *KIT* duplication breakpoints among pig breeds.

	Distance from duplication breakpoint[a]											
5′ breakpoint	−334	−98	10	46	209	376	412	445	474	502	1697	2653
832E11[b]	-[c]	A	C	C	T	A	C	C	T	A	A_{11}	C
953F11[b]			-	-	-	-	-	-	-	-	A_{10}	T
Wild boar	-	-	G	-	-	-	-	-	-	-		
Hampshire	-	-	G	-	-	-	-	-	-	-		
Duroc	-	-	G	-	-	-	-	-	-	-		
Meishan	N_{16}	G	G	A	G	G	T	T	C	C		
Piétrain	-	-	G	-	-	-	-	-	-	-		
Large White	-	-	G	-	-	-	-	-	-	-		
Landrace	-	-	-/G	-	-	-	-	-	-	-		

	Distance from duplication breakpoint[a]								
3′ breakpoint	−213	−168	−52	40	82	297	312	401	442
1041B3[b]	C	C	T	C	C	G	A	C	T
953F11[b]	-	-	-						
Wild boar	-	-	-	-	-	-	-	-	-
Hampshire	-	-	-	-	-	-	-	-	-
Duroc	-	T	C	T	T	T	G	T	G
Meishan	T	-	-	T	-	-	-	-	-
Piétrain	-	-	-	-	-	-	-	-	-
Large White	-	-/T	-/C	-/T	-/T	-/T	-/G	-/T	-/G
Landrace	-	-/T	-/C	-/T	-/T	-/T	-/G	-/T	-/G

[a] A negative value indicates that the position is located 5′ of the actual duplication breakpoint according to the orientation given in Fig. 2.
[b] BAC clones.
[c] A dash indicates identity to the master sequence; -/G, etc. indicates heterozygosity; blank indicates that the position is not present (BAC 953F11) or that the position was not investigated with genomic DNA.

Table 3. Presence of the unique *KIT* duplication breakpoint associated with Dominant White color in different breeds of pigs.

Population	Coat color	Presumed genotype[a]	Presence of duplication		
			+	−	Total
Wild boar	Wild type	i/i	0	2	2
Large White	White	$I^?/I^?$	11	0	11
Landrace	White	$I^?/I^?$	4	0	4
Piétrain	White/black spots	i/i	4	1	5
Berkshire	Black/white points	i/i	0	7	7
Duroc	Red	i/i	0	4	4
Hampshire	Black/white belt	I^{Be}/I^{Be}	0	4	4
Linderöd	Red/black spots	i/i	0	1	1
Meishan	Black	i/i	0	3	3

[a] $I^?$ indicates that the allele may be I or I^P, both carrying the *KIT* duplication.

substitution (Table 2). The results indicate that the *KIT* duplication occurred recently or that the sequences have been homogenized by gene conversion.

Distribution of the KIT duplication among pig breeds. Southern blot analysis of *Hind*III-digested genomic DNA was used to exclude the possibility that BAC 953F11 was a cloning artifact. A PCR fragment free of repetitive sequences from subclone pUC953H4 was used as probe. The results showed that the unique 3.9-kb *Hind*III fragment was present in Large White and Landrace animals carrying various *Dominant White* alleles but not in the Wild boar (Fig. 3B). A considerable variation in the hybridization signal of the 3.9-kb fragment was observed, consistent with our recent observation of a variation in *KIT* copy number in Large White and Landrace pigs using a quantitative test for the splice mutation (Pielberg et al. 2002).

PCR screening showed that the *KIT* duplication was present in Large White and Landrace animals but not in Wild boar, Berkshire, Duroc, Hampshire, Linderöd, or Meishan pigs as expected (Table 3). Unexpectedly, four out of five Piétrain pigs were positive for the duplication. Interestingly, the single Piétrain pig that did not carry the duplication was atypical for the breed and almost entirely black. Its parents had the usual color and carried the duplication. All Piétrain pigs with the duplication were negative for the splice mutation, indicating that they carry the I^P allele.

A fragment of 864 bp from the 3′-5′ duplication breakpoint region (present in BAC 953F11) was PCR amplified from Large White and Landrace animals carrying the I allele and from Large White and Piétrain animals carrying the I^P allele. The sequences obtained from 10 animals were completely identical, demonstrating that the I and I^P alleles originate from the same duplication event. A similar sequence comparison across breeds was carried out for 968 bp from the 5′ duplication breakpoint (BAC 832E11) and for 1022 bp from the 3′ duplication breakpoint (BAC 1041B3). The sequence of the corresponding region in 953F11 was almost completely identical to the sequence obtained from the European Wild Boar and several European domestic pig breeds. In contrast, the sequence from the Chinese Meishan breed differed by 10 single nucleotide substitutions and one 16-bp insertion/deletion from the sequence of the European Wild Boar. Two divergent haplotypes of the region corresponding to the 3′ duplication breakpoint (BAC 1041B3) were observed among European domestic pigs. The two haplotypes differed by as many as eight substitutions within a region of about 600 nucleotides (Table 2). The two Landrace and Large White animals tested were both heterozygous for these divergent haplotypes.

Table 4. Gene-specific PCR primers and Sequence Tagged Sites (STS) from the *PDGFRA–KTT–KDR* region in pigs.

Gene/STS name	Forward Primer 5′-3′	Reverse Primer 5′-3′
211E123′	aatgttgcaggattcaccgtcac	gctctgagcccttggatcatg
211E125′	gaacatggcctcagctcttgg	agaagagactgcagctcgtaccc
280B33′	tgcatttctattcagcaccgg	ggatgggtttcagaaagggc
280B35′	aaccagcagtcatttgagcca	ggatcagctcaagaaaaggca
280D75′	ctccaccacagacagttcttctc	caggaagcccaaaggatatacac
311F43′	ctggttatggacggtggctgag	ggttagatgatgaagagtgggtg
311F45′	tgagtagtgggagtggcgggtg	gccgagcaaagatgaagagaggtc
513A23′	aatgcacatggcagctcacc	caaacagggcactggtcaacc
513A25′	aaggactcatgggcttggg	catcagaggtttctttgcagtgg
549C33′	acatgcttcctcaagtagttc	tccagagacaatagtgcagc
549C35′	gaagtcagcaagacagggcc	gcagatcagagcattccccac
645D53′	gaagagttctagagtattctccg	tgtttgctttgccctctgtatg
717E83′	ctgggcattctgtgctaagg	agtcagattcatttccactgc
717E85′	cttattccactcccagaacc	tgcttcgccaggcaaaggag
723G23′	ccttggtgcacagctgaactt	cagcaggatgaaggagcacac
723G25′	cccctgtcttatggatactagtca	gaagcatgcattctattccga
757A83′	gctccgagcctccagcacac	cagggatcgaacccacaacc
763F13′	tgcaggttcataaagccgaga	ttcctaccacacatggcagct
798C55′	cttattccactcccagaacc	tgcttcgccaggcaaaggag
832E113′	aatcctccctcatataagttcccg	tgctctccaccaaggctgaagg
832E115′	tccccaagcctctattctgtagtc	tttgatcacatgccctgcac
927C35′	ttactgcaatcccctgcctc	gcactcagaaatggtttgtttttc
953F113′	agcctgaaccaccccattg	tgcacagggaggagtcctgt
953F115′	tccagggcactctttgatctg	agggaagcaggaaaagaccac
1000D23′	ttatggacactctggagccc	gcctgctaacaattatgcatg
1000D25′	aaaagcattctgtcctctcctg	tagcactgcaccatgaagggaac
1041B33′	tcctcctgcgtctctgacag	tggctgaccctgaacatctc
1044B123′	gcttcactatcaatattccagc	ttagtttagtctcactttgtgc
KIT, exon 1	ggctctgggggctcggctttgc	tgtctggacgcgaagcaagagctg
KIT, exon 17	gtattcacagagacttggcggc	ggggctgcagttcgctaagttg
KIT, exon 2	tcaaccatctgtgagtccagggg	tgttggtgcacgtgtatttgcc
KIT, exon 6	aatgatggcgaggatgtggatc	tctcagacttgggataatcctcc
PDGFRA, exon 22	ctggacttcctraagagtgacc	tgytcrtccagrccaccytccc
PDGFRA, exon 23	gacrggttccagyagttccacc	gaagctgtcytccaccaggtc

$y = c/t$, $r = a/g$

Discussion

This study demonstrates the power of comparative genomics and is an advance in pig coat color genetics. We used human map data during the construction of the BAC contig and comparative mouse data when interpreting the phenotypic effects of the *KIT* mutations. We show that the *Patch* (I^P) and *Dominant White* (*I*) alleles carry the same duplication. The most likely evolutionary history is thus that the duplication arose first, causing a partial dominant white phenotype, followed by the occurrence of the splice mutation in one of the copies, causing the fully dominant white phenotype. As expected, the *KIT* duplication breakpoint was found in Large White and Landrace pigs but not in colored breeds. However, an unexpected finding was that Piétrain pigs carry the *KIT* duplication and the I^P allele. Piétrain pigs have a white coat with black spots and have previously been assumed to have the wild type (*i*/*i*) genotype (Legault 1998). The origin of the Piétrain breed is somewhat obscure, but it has been proposed that the Large White were one of the breeds used during the development of the Piétrain (Jones 1998), which may explain the presence of the *KIT* duplication. Piétrain and Berkshire pigs are both homozygous for the E^P allele at the *E*/*MC1R* locus, causing a black-spotted phenotype, but Berkshire is almost solid black (Kijas et al. 2001). Sewall Wright proposed as early as 1918 that the black color of Berkshire is an extended form of black spotting (Wright 1918). This study indicates that *KIT* is a major modifying locus for the extension of black spotting.

The size of the duplication was estimated at about 450 kb with the BAC contig. The duplication appears to be a "recent" event that most likely occurred subsequent to domestication. The restriction fragment analysis of BAC clones as well as the sequence analysis of the breakpoint region showed that the two copies are almost identical. However, the two copies may show concerted evolution due to the occurrence of unequal crossing-over and gene conversion. We have recently shown that the number of *KIT* copies per chromosome varies between one and three in white breeds (Pielberg et al. 2002). The variation in copy number is most likely generated by unequal crossing-over. The observation of an almost black Piétrain pig without the *KIT* duplication indicates that a similar variability in *KIT* copies also occurs in Piétrain pigs.

Pigs have been domesticated from distinct subspecies of Wild Boars in Europe and Asia (Giuffra et al. 2000). Furthermore, Asian domestic pigs were introgressed into European breeds during the 18th and 19th centuries, and the white breeds Large White and Landrace both have a considerable proportion of Asian ancestry. The results in this study allowed us to investigate whether the *KIT* duplication and the Dominant White color first arose in Europe or Asia. The sequence of the unique duplication breakpoint supports a European origin, since it was almost identical to the corresponding sequences on non-duplicated chromosomes from the European Wild Boar but showed multiple differences to sequences from the single Asian breed (Meishan) investigated in this study.

Repetitive elements are highly abundant in mammalian genomes. Bioinformatic analysis of the human draft genome sequence shows that about 45% of the human genome is derived from transposable elements (International Consortium 2001). There are several examples of human disorders in which the causative mutation constitutes an insertion of a SINE or LINE element (Kazazian and Moran 1998), and a deletion in the beige gene in the rat occurred by homologous recombination between two LINE1 elements (Mori et al. 2001). Retrotransposons may also act as a creative force by generating new genes or altering the regulation of existing genes (Kazazian 2000). As an example, the duplication of the γ-globin genes present in most primates originated by unequal

crossing-over between L1 elements (Fitch et al. 1991). This study provides another example of an evolutionary novelty created by a recombination event between two LINE elements leading to a gene duplication with a phenotypic effect. The duplication does not appear to have any major deleterious effects, since millions of highly productive domestic pigs around the world are homozygous for the duplication.

The order of the three protein tyrosine kinase genes *PDGFRA–KIT–KDR* is conserved between human Chr 4 and pig Chr 8. The three genes are organized head-to-tail in humans (Spritz et al. 1994), and we have shown that this is the case also for *KIT* and *PDGFRA* in pigs. The tandem duplication involves the entire coding sequence of *KIT*, but not *PDGFRA* and *KDR*. The location of the duplication breakpoint in the 5′ upstream region of *KIT* suggests a reasonable explanation why the duplication causes a distinct phenotypic effect. Chromosomal rearrangements in this region in the mouse are associated with ectopic expression of *KIT* and dominant pigmentation disorders (Berrozpe et al. 1999; Hough et al. 1998). In fact, comparative sequence analysis shows that the inversion breakpoint associated with *Rump-white* and the deletion breakpoint associated with *Patch* in the mouse are both located about 30 kb further upstream of *KIT* than the pig duplication breakpoint. Thus, the comparative data strongly suggest that the duplicated copy in pigs is dysregulated, since it does not contain all regulatory elements needed for normal expression.

This BAC contig will be an excellent resource for comparative sequencing of the region between *PDGFRA* and *KIT*. This is probably the most powerful approach for finding the tissue-specific control elements upstream of *KIT*, postulated on the basis of the observed association between dominant *KIT* mutations and chromosomal rearrangements far upstream of the coding sequence in pigs and mice (Berrozpe et al. 1999; Hough et al. 1998). In the present study, we identified an Evolutionary Conserved Region of about 1 kb between pigs and human located close to one of the duplication breakpoints (Fig. 4A). The sequences do not contain any obvious open reading frames and do not show any significant sequence similarity to other sequences in public databases. We therefore assume that this conserved region represents a regulatory element. The striking similarity in the content and orientation of retrotransposons in this region between pig and human (Fig. 4A) implies that the integrations of the L1 and MER58B elements predate their divergence from a common ancestor. Furthermore, the surprisingly high sequence similarity between the repeats in the two species suggests that they may have a functional role and may have been maintained by selection. Accumulating data suggest that the integration of LINE elements may influence gene regulation (Yang et al. 1998).

Acknowledgments. Sincere thanks are due to Ulla Gustafsson and Signe Hässler for valuable assistance; to Lena Claesson-Welsh, who kindly provided the human *KDR* cDNA clone; and to Göran Andersson and Graham Plastow for discussions. The Pig Improvement Company (PIC) provided samples from Piétrain and Berkshire pigs, and funded the study together with the Swedish Research Council for Forestry and Agriculture.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215, 403–410

Berrozpe G, Timokhina I, Yukl S, Tajima Y, Ono M et al. (1999) The W^{sh}, W^{57}, and *Ph Kit* expression mutations define tissue-specific control elements located between −23 and –154 kb upstream of *Kit*. Blood 94, 2658–2666

Chowdhary BP, de la Sena C, Harbitz I, Eriksson L, Gustavsson I (1995) FISH on metaphase and interphase chromosomes demonstrates the physical order of the genes for GPI, CRC, and LIPE in pigs. Cytogenet Cell Genet 71, 175–178

Claesson-Welsh L, Eriksson A, Westermark B, Heldin CH (1989) cDNA cloning and expression of the human A-type platelet-derived growth factor (PDGF) receptor establishes structural similarity to the B-type PDGF receptor. Proc Natl Acad Sci USA 86, 4917–4921

Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L et al. (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc Natl Acad Sci USA 88, 7396–7400

Frengen E, Thomsen P, Kristensen T, Kran S, Miller R et al. (1991) Porcine SINEs: characterization and use in species-specific amplification. Genomics 10, 949–956

Giuffra E, Evans G, Törnsten A, Wales R, Day A et al. (1999) The Belt mutation in pigs is an allele at the *Dominant white* (*I*/*KIT*) locus. Mamm Genome 10, 1132–1136

Giuffra E, Kijas JMH, Amarger V, Carlborg Ö, Jeon J-T et al. (2000) The origin of the domestic pig: independent domestication and subsequent introgression. Genetics 154, 1785–1791

Hough RB, Lengeling A, Bedian V, Lo C, Bucan M (1998) *Rump white* inversion in the mouse disrupts dipeptidyl aminopeptidase-like protein 6 and causes dysregulation of *Kit* expression. Proc Natl Acad Sci USA 95,13800–13805

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

Jeon J-T, Amarger V, Robic A, Rogel-Gaillard C, Bongcam-Rudloff E, et al. (2001) Comparative analysis of a BAC contig of the porcine *RN* region and the human transcript map: implications for the cloning of trait loci. Genomics 72, 297–303

Johansson Moller M, Chaudhary R, Hellmen E, Hoyheim B, Chowdhary B et al. (1996) Pigs with the dominant white coat color phenotype carry a duplication of the *KIT* gene encoding the mast/stem cell growth factor receptor. Mamm Genome 7, 822–830

Jones GF (1998) Genetic aspects of domestication, common breeds and their origin. In: *The Genetics of the Pig*, A Ruvinsky, MF Rothschild, eds. (Oxon, UK: CAB International), pp 17–50

Kaplan DJ, Jurka J, Solus JF, Duncan CH (1991) Medium reiteration frequency repetitive sequences in the human genome. Nucleic Acids Res 17, 4731–4738

Kazazian Jr, HH (2000) L1 retrotransposons shape the mammalian genome. Science 289, 1152–1153

Kazazian Jr, HH, Moran JV (1998) The impact of L1 retrotransposons on the human genome. Nat Genet 19, 19–24

Kijas JMH, 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, 779–785

Legault C (1998) Genetics of colour variation. In: *The Genetics of the pig*, MF Rothschild, A Ruvinsky, eds. (Oxon, UK: CAB International), pp 51–69

Marklund S, Kijas J, Rodriguez-Martinez H, Ronnstrand L, Funa K et al. (1998) Molecular basis for the dominant white phenotype in the domestic pig. Genome Res 8, 826–833

Marklund S, Moller M, Sandberg K, Andersson L (1999) Close association between sequence polymorphism in the *KIT* gene and the roan coat color in horses. Mamm Genome 10, 283–288

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM et al. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16, 1046–1047

Mori M, Nishikawa T, Higuchi K, Nishimura M (2001) Deletion in the beige gene of the beige rat owing to recombination between LINE1s. Mamm Genome 10, 692–695

Pielberg G, Olsson C, Syvänen A-C, Andersson L (2002) Unexpectedly high allelic diversity at the *KIT* locus causing dominant white color in the domestic pig. Genetics 160, 305–311

Reinsch N, Thomsen H, Xu N, Brink M, Looft C et al. (1999) A QTL for the degree of spotting in cattle shows synteny with the KIT locus on chromosome 6. J Hered 90, 629–634

Rogel-Gaillard C, Billault A, Bourgeaux N, Vaiman M, Chardon P (1999) Characterisation and mapping of type C endogenous retroviral element in swine using a BAC library. Cytogenet Cell Genet 85, 273–278

Smit AFA, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246, 401–417

Spritz RA, Strunk KM, Lee ST, Lu-Kuo JM, Ward DC et al. (1994) A YAC contig spanning a cluster of human type III receptor protein tyrosine kinase genes (*PDGFRA–KIT–KDR*) in chromosome segment 4q12. Genomics 22, 431–436

Wright S (1918) Color inheritance in mammals. VIII. Swine. J Hered 9, 33–38

Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R (1998) Apolipoprotein(a) gene enhancer resides within a LINE element. J Biol Chem 273, 891–897

Website References

Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov
MGI, http://www.informatics.jax.org
Repeatmasker, http://ftp.genome.washington.edu/cgi-bin/RepeatMasker
NCBI Blast server, http://www.ncbi.nlm.nih.gov
ENSEML, http://www.enseml.org
VISTA, http://www-gsd.lbl.gov