SUPPLEMENTARY INFORMATION

doi:10.1038/nature11622

1. Genome assembly

Sequencing

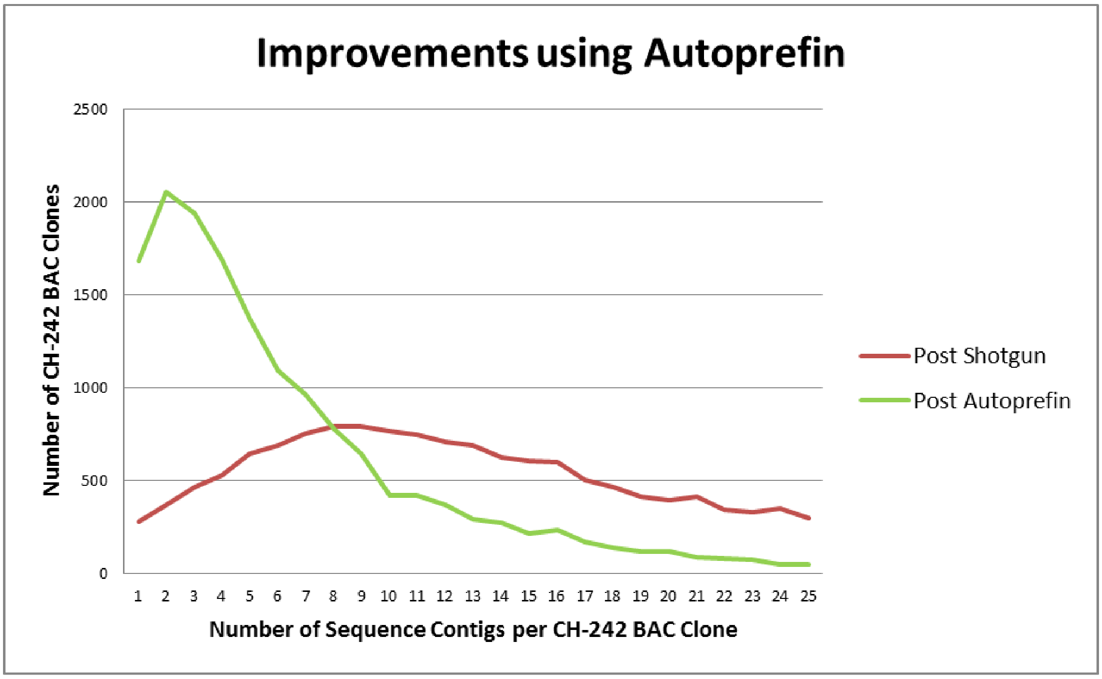
The pig genome was sequenced following a hybrid approach representing a refinement of the strategy announced earlier (Schook et al 2005). The sequence predominantly represents the genome of a single female from the Duroc breed (Duroc 2-14), but some regions, including self-evidently the Y chromosome, are derived from the genomes of other pigs that are represented in bacterial artificial chromosome (BAC) libraries.

Clone based sequencing

Briefly, BAC and fosmid clones selected from a minimal tile path across the genome were identified from the high resolution physical (BAC contig) map (Humphray et al 2007) and were subjected to hierarchical shotgun sequencing. The minimal tile path (MTP) provides coverage of 98.3% of this physical map. The contig map was constructed from clones from four BAC libraries: 1) CHORI-242 (http://bacpac.chori.org/porcine242.htm/) generated from a single Duroc sow (i.e. Duroc 2-14); 2) PigEBAC generated from a single male Large White x Meishan F1) (Anderson et al 2000); 3) RPCI-44 (constructed from pooled material from 4 male pigs, 1/4 Meishan, 3/8 Yorkshire, and 3/8 Landrace) (Farhernkrug et al. 2001) and 4) the INRA BAC library (made from a single Large White male) (Rogel-Gaillard et al 1999). In addition, a fosmid library was constructed from Duroc 2-14 DNA and 565,349 fosmid end sequences were generated by single pass sequencing (Accession numbers: HE000001 to HE565349). BAC clones from the CHORI-242 library were preferentially chosen for sequencing. Clones from the MTP were selected for sequencing following an iterative approach. In order to identify clones that closed gaps in the map whilst minimising the overlaps, BAC and fosmid end sequences were aligned against the already sequenced clones. Where there was no suitable CHORI-242 BAC clone, clones from other BAC libraries were selected for sequencing – sequences from ~200 such clones from the PigE BAC library were included representing a ~1.3% contribution to the current assembly. The fosmid library was used for map and sequence closure in the later stages of the project.

For each BAC selected for sequencing paired-end reads were generated for 768 subclones using Sanger capillary technology (average read length of 707 base pairs (bp)) to provide ~4x sequence coverage. Most BAC clones were subsequently subjected to one round of automated pre-finishing by primer walking from the ends of the sequence contigs constructed from the initial 4x coverage. This hierarchical shotgun sequencing was primarily undertaken at the Wellcome Trust Sanger Institute (WTSI), with additional clones sequenced by the National Institute of Agrobiological Sciences, Japan.

After the shotgun process and prior to automated pre-finishing, 279 Pig clones were contiguous, 2,293 in 5 contigs or less and 6,091 in 10 contigs or less. Following pre-finishing, 1,681 Pig clones were contiguous, 8,742 in 5 contigs or less and 12,674 in 10 contigs or less (see Supplementary Fig. 1). The increased sequence contiguity achieved was equivalent to the addition of 2x sequence coverage and was achieved at significantly lower cost.



Supplementary Fig. 1: Number of sequence contigs per sequenced BAC before and after auto-prefinishing.

Whole genome shotgun (WGS) sequencing

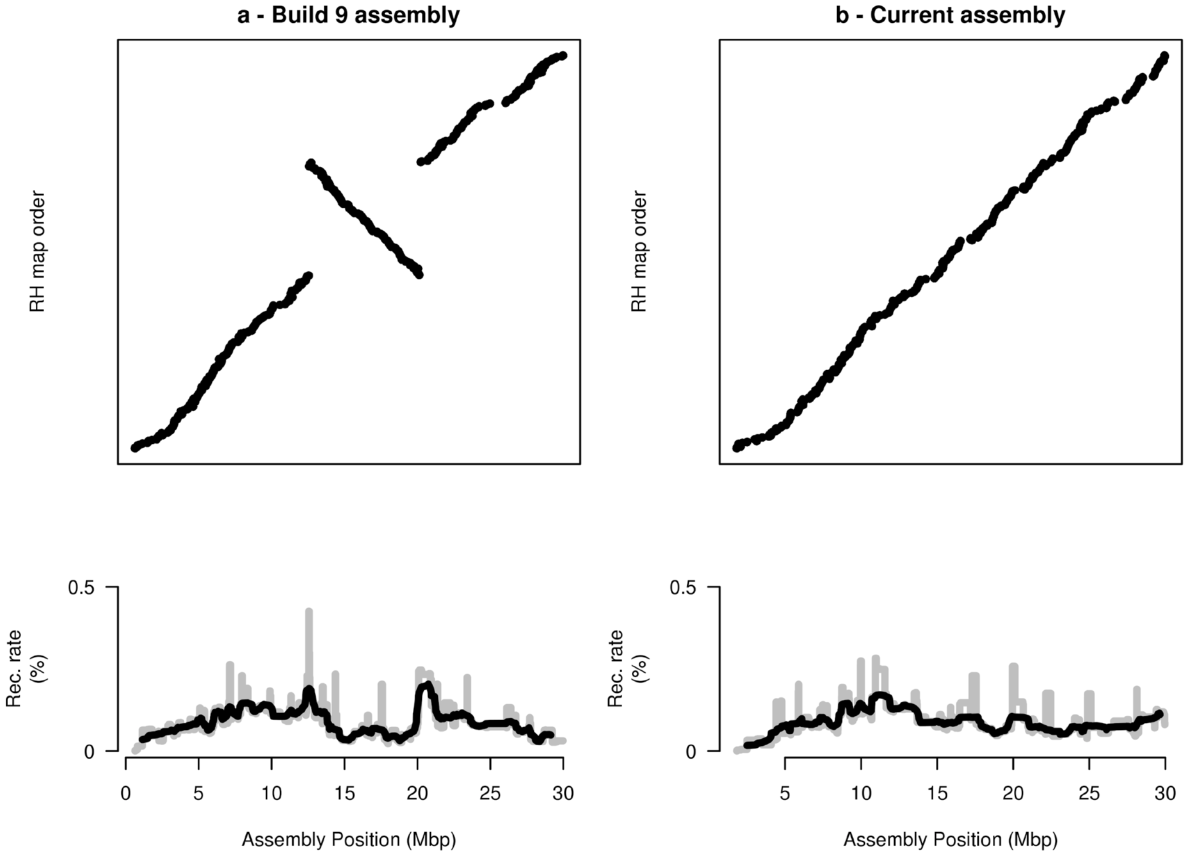
Clone-based sequence coverage was supplemented with whole genome shotgun (WGS) sequence data generated from DNA isolated from the same animal (Duroc 2-14). These WGS data were generated using Illumina/Solexa sequencing at BGI (formerly known as the Beijing Genomics Institute) and the Wellcome Trust Sanger Institute (WTSI). The Illumina short reads comprised 66.5 x 10^9 base pairs (Gb) of paired-end 44 bp reads (BGI) and 40 Gb of paired-end 108 bp reads (WTSI).

Genome assembly

The BAC-derived sequence data were assembled into sequence contigs using Phrap (http://www.phrap.org/) on a clone-by-clone basis as the data were generated and the resulting sequences deposited into the ENA/Genbank/DDBJ public databases. After checking the order and orientation of the sequence contigs (see below), the overlapping sequenced BAC clones were assembled into contigs and scaffolds to cover each chromosome using SSAHA (http://www.sanger.ac.uk/resources/software/ssaha/, Ning et al 2001).

Chromosome assignment, order and orientation

Chromosome assignment and long range order and orientation of scaffolds were checked by alignment with the clone-based physical map (Humphray et al., 2007) and recombination and radiation hybrid (RH) maps (Tortereau et al 2012, Servin et al. 2012). With the exception of an unresolved conflict between the sequence assembly and the RH map on SSC 2p, the sequence and RH maps are essentially co-linear. Thus, we conclude that the chromosome assignments and long range order and orientation of the scaffolds are robust. The use of alignments with the RH maps to confirm order and orientation is illustrated in Supplementary Figure 2.



Supplementary Figure 2: This figure presents the first 30 Megabases of pig chromosome 6 and its comparison with the RH map, for the build 9 assembly (left) and the current pig genome Sscrofa10.2 assembly (right) . A pointwise (gray) and smoothed (black) estimate of the recombination rate along the assembly is shown for the same region. The figure clearly shows that resolving the inversion in the previous build reduced the recombination rate in the region, and in particular two recombination spikes at the border of the inversion. For further details see Servin et al. 2012.

The BAC clones from which the backbone of the draft genome sequence was constructed provide 15,543 bins into which the clone-derived sequence contigs could be robustly assigned (Supplementary Table 1). The order and orientation of sequence contigs within each BAC was informed by knowledge of a) BAC end sequences, b) paired-end sequences of the sub-clones and c) gene models. The order and orientation of some contigs were also resolved through incorporation of the WGS contigs. This information was used to group, order and orient sequence contigs into fragment chains. However, the order and orientation of some sequence contigs within the BAC clone bins remains unresolved.

Supplementary Table 1. Summary of sequenced BAC clones

Chromosome	Unique Clones	Finished Clones	Draft Clones
1	1,882	40	1,842
2	966	111	855
3	860	7	853
4	885	56	829
5	681	10	671
6	909	43	866
7	826	131	695
8	870	2	868
9	931	6	925
10	473	0	473
11	519	6	513
12	386	22	364
13	1,264	6	1,258
14	963	38	925
15	924	2	922
16	521	4	517
17	445	70	375
18	373	20	353
X	854	379	475
Y	11	0	11
Total	15,543	953	14,590

Integration of WGS data

The WGS short reads were independently assembled using a) SOAPdenovo (Li et al. 2010) and b) Cortex (Iqbal et al. 2012). The resulting sequence contigs were aligned with the BAC-derived contigs using the alignment tool BLAT (Kent 2002). Consequently, the WGS contigs were used to extend BAC clone-derived sequence contigs, close gaps between clone-derived contigs and where the WGS contigs and scaffolds did not align with the BAC-derived contigs they were identified as new 'unplaced' scaffolds which, as yet, are not anchored to chromosomes. As the ends of contigs are more error prone, the ends of each BAC contig were trimmed by 50-100 bp (depending on the alignment profiles) prior to searching for matches with the WGS contigs. About 3,000 gaps in the BAC contigs were closed with Cortex and SOAPdenovo WGS contigs. The WGS contigs/scaffolds not only facilitated the closure of gaps, but also the resolution of order and orientation for some BAC contigs. Finally, the new unassigned WGS scaffolds comprise ~212 Mbp and contain interesting genes including CD163 and IGF2.

Assembly statistics and quality

The summary statistics for the draft pig genome sequence (Sscrofa10.2) are presented in Supplementary Table 2. The terms contigs and scaffolds are used with their original meanings: contigs = continuous sequence with no stretch of unknown bases (Ns); scaffold are sequences consisting of contigs linked, ordered and oriented through paired-end sequences, but with gaps of unknown bases and size.

The assembly of genome sequence data from whole genome shotgun sequences generated with next-generation sequencing technologies has required a reassessment of the standards applied to genome sequences (Chain et al. 2009). The Sscrofa10.2 assembly meets the "***Improved High-Quality Draft***" classification proposed by Chain et al. (2009) on several grounds. Both automated and manual methods were used to improve the sequence after the initial shotgun sequencing of the BAC clones. The improvements in contiguity achieved through the automated prefinishing are evident in the consequent reduction in the number of contigs per BAC clone (Supplementary Figure 1). Furthermore 953 of the BAC clones were manually improved to "finished" quality.

Comparisons of the Sscrofa10.2 assembly with draft genome sequences from other species confirm the merits of the assembly. Contiguity, for which contig N50 is regarded as the diagnostic statistic, is a key measure of the quality of a sequence assembly. The contig N50 of 80,720 bp and 69,669 bp for the sequence assigned to chromosomes and the global measure for the assembly, respectively, (Supplementary Table 3) compare favourably with those of the published draft cattle genome (Bovine Genome Sequencing and Analysis Consortium, 2009, Btau3.1) (76,449 bp) but less so with the published draft horse genome (Wade et al 2009) (112,381 bp). The horse genome is a high quality assembly of solely WGS Sanger paired end reads from a range of cloned inserts of plasmids, fosmids and BACs. The use of paired end short reads from next-generation sequencing technologies yield assemblies with significantly lower contiguity, for example the contig N50 for the panda genome is 39,886 bp. The recently published gorilla genome sequence (Scally et al 2012) which combined low coverage WGS Sanger reads with high coverage Illumina WGS reads yielded an assembly with a contig N50 of 11,661 bp and confirmed the difficulties of hybrid assembly construction.

The order of the sequenced BAC clones in the physical map (Humphray et al 2007) was used for scaffolding rather than information from long range paired reads such as BAC or fosmid end sequences. The alignments with the radiation hybrid and linkage maps (see above and Tortereau et al 2012, Servin et al. 2012) demonstrate the high quality of the medium to long range order and orientation of the assembly.

A key strength of this draft pig genome sequence is the associated large fragment clone resource. The BAC clones, from which most of the sequence assigned to chromosomes was generated, are available for targeted sequence improvement and closure. Moreover, the BAC physical map (Humphray et al 2007; http://pre.ensembl.org/Sus_scrofa_map/Info/Index) and the alignment of BAC end and fosmid end sequences with the genome sequence allows the selection of clones for targeted gap and sequence closure. Scientists at the Wellcome Trust Sanger Institute and University of Cambridge are currently exploiting these resources to establish a 'finished' quality sequence of the pig chromosome X.

The final assembly (Sscrofa10.2) has been deposited in the public sequence databases [Accession number: AEMK01000000]. The primary source of the Sscrofa10.2 assembly is the NCBI ftp site:

ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Sus_scrofa/Sscrofa10.2/

The chromosomes are CM000812-CM00830 and CM001155. They are built from 5,343 placed scaffolds whose accessions are in the ranges GL878569-GL882503 and JH114391-JH118402. The 4,562 unplaced scaffolds of Sscrofa10.2 have accessions in the ranges GL892100-GL896682 and JH118403-JH118999.

Supplementary Table 2. Summary statistics for the draft pig genome sequence build 10.2

Chromosome	Total length	Ungapped length	Number of scaffolds	Scaffold N50	Number of contigs	WGS contigs	Spanned gaps	Unspanned gaps
1	315,321,322	279,914,422	692	562,366	9,261	426	8,569	691
2	162,569,375	145,631,175	331	692,939	4,713	148	4,382	330
3	144,787,322	129,207,922	304	654,307	4,598	170	4,294	303
4	143,465,943	129,362,743	276	715,815	3,808	148	3,532	275
5	111,506,441	99,275,441	239	618,165	3,549	138	3,310	238
6	157,765,593	139,069,593	366	529,884	4,826	150	4,460	365
7	134,764,511	121,271,311	266	712,505	2,698	120	2,432	265
8	148,491,826	132,692,726	309	573,231	4,300	166	3,991	308
9	153,670,197	139,490,897	276	713,366	4,569	187	4,293	275
10	79,102,373	71,122,273	156	628,080	2,457	76	2,301	155
11	87,690,581	77,960,681	190	555,481	2,989	97	2,799	189
12	63,588,571	56,400,871	141	543,619	2,018	52	1,877	140
13	218,635,234	195,589,234	449	629,709	6,909	260	6,460	448
14	153,851,969	140,665,969	259	807,032	3,119	125	2,860	258
15	157,681,621	140,675,921	332	617,435	4,889	227	4,557	331
16	86,898,991	78,720,191	160	743,205	2,448	88	2,288	159
17	69,701,581	62,138,581	148	638,723	2,278	61	2,130	147
18	61,220,071	55,640,371	109	813,903	1,906	80	1,797	108
X	144,288,218	127,507,118	333	584,790	2,144	40	1,811	332
Y	1,637,716	1,333,916	7	207,906	45	0	38	6
Totals (placed)	2,596,639,456	2,323,671,356	5,343		73,524	2,759	68,181	5,323
Unplaced scaffolds	211,869,922	195,490,322	4,562	98,022	168,358	168,358	163,796	0

N50: 50% of the genome / chromosome is in scaffolds of this length or longer.

Whilst the clone-derived, chromosome-assigned draft sequence has a high degree of contiguity, the WGS unassigned parts of the assembly are highly fragmented. In order to reflect this differentiation, the standard quality measures (average contig and scaffold sizes; N50, etc..) are presented separately for the chromosome assemblies and Chr U (Supplementary Table 3). For completeness the summary statistics for both placed and unplaced scaffolds are also listed. However, given the differences in the nature of the sequence data (predominantly Sanger reads of ~700 bp versus Illumina reads of 44-100 bp) and the assembly procedures these global statistics should be used with caution.

Supplementary Table 3. Summary standard quality measures for porcine genome build 10.2

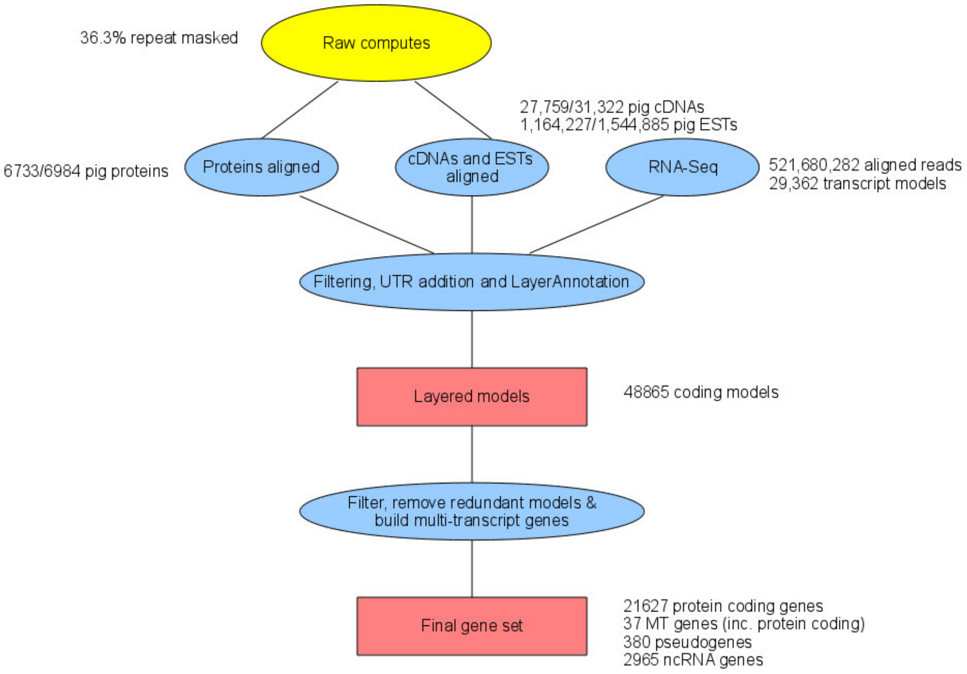
	Length (bp)
Chromosomes 1-18, X, Y	
Contigs N50	80,720
Contigs N90	13,487
Average contig length	31,604
Largest contig length	1,598,650
Scaffold N50	637,332
Scaffold N90	189,449
Average scaffold length	436,176
Largest scaffold length	3,862,550
Chr U (unplaced scaffolds)	
Contigs N50	2,423
Contigs N90	583
Average contig length	1,161
Largest contig length	22,846
Scaffold N50	98,022
Scaffold N90	27,533
Average scaffold length	46,442
Largest scaffold length	59,4937
Total assembly (placed + unplaced scaffolds)	
Contigs N50	69,669
Average contig length	11,611
Largest contig length	1,598,650
Scaffold N50	576,008
Average scaffold length	283,544
Largest scaffold length	3,862,550

N50: 50% of the genome in in fragments of this length or longer

2. Annotation and Ensembl Gene Build

Raw Computes Stage: Searching for sequence patterns, aligning proteins and cDNAs to the genome.

The annotation process of the Sscrofa10.2 assembly began with the “raw compute” stage (Supplementary Fig. 3) whereby the genomic sequence was screened for sequence patterns including repeats using RepeatMasker (Smit et al) (version 3.2.8 with parameters '-nolow -s -species pig'), Dust (Kuzio et al 2006) and TRF (Benson 1999). RepeatMasker masked 36.3% of the genome. Adding low complexity masking (including gaps) with Dust brings the total masked to 48.2%.



Supplementary Fig. 3: Summary of pig gene annotation project.

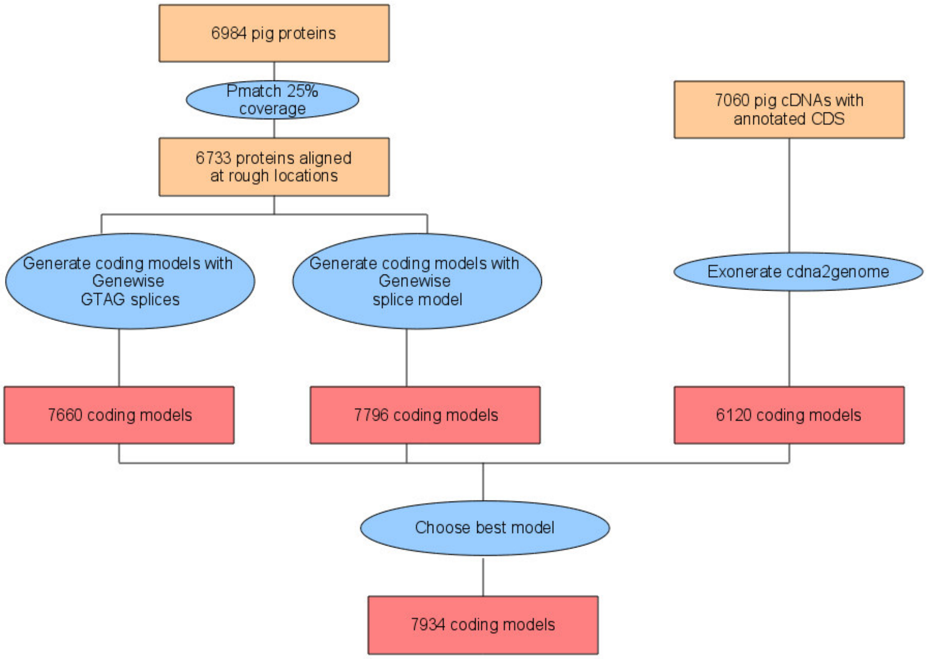
Transcription start sites were predicted using Eponine-scan (Down, Hubbard 2002), and FirstEF (Davuluri et al 2001) CpG islands and tRNAs (Lowe et al 1997) were also predicted.

Genscan (Burge, Karlin 1997) was run across RepeatMasked sequence and the results were used as input for UniProt (Goujon et al 2010), UniGene (Sayers et al 2010), and Vertebrate RNA (ENA) alignments by WU-BLAST (Altschul et al 1990) (passing only Genscan results to BLAST is an effective way of reducing the search space and therefore the computational resources required). This resulted in 287,736 UniProt vertebrate and

72,840 UniProt non-vertebrate sequences, 340,390 UniGene and 327,557 Vertebrate RNA sequences aligning to the genome.

Targeted Stage: Generating coding models from pig evidence

Pig protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL (Goujon et al 2010) and Genbank) and filtered to remove sequences based on predictions. The pig sequences were mapped to the genome using Pmatch as indicated in Supplementary Fig. 4. For pig cDNA sequences with annotated CDS, Exonerate's (Slater, Birney 2005) cdna2genome model was used.


6984 pig proteins
Pmatch 25% coverage
6733 proteins aligned at rough locations
7060 pig cDNAs with annotated CDS
Generate coding models with Genewise GTAG splices
Generate coding models with Genewise splice model
Exonerate cdna2genome
7660 coding models
7796 coding models
6120 coding models
Choose best model
7934 coding models

Supplementary Fig. 4: Targeted stage using pig specific proteins and cDNAs with annotated CDS

Models of the coding sequence (CDS) were produced from the proteins using Genewise (Eyras et al 2004). Two sets of models were produced: one with only consensus splice sites and one where non-consensus splices were allowed.

Where a single protein sequence generated two different coding models at the same locus, the BestTargetted module was used to select the coding model that most closely matched the source protein to take through to the next stage of the gene annotation process. The generation of transcript models using pig-specific data is referred to as the "Targeted stage". This stage resulted in 7,934 coding models.

cDNA, EST and ENSSSCP alignment

Pig cDNAs were downloaded from Genbank, clipped to remove polyA tails, and aligned to the genome using Exonerate. Of these, 27,759 of 31,322 pig cDNAs aligned with a cut-off of 50% coverage and 50% identity. Expressed Sequence Tags (ESTs) were downloaded from the same sources and processed in the same way. Of these, 1,164,227 of 1,544,885 pig ESTs aligned with a cut-off of 80% coverage and 90% identity.

ENSSSCP models from Ensembl 64 were also aligned to the genome. 17,493 canonical translations were downloaded and aligned using Exonerate. Of these 16,399 aligned with identity greater than 95% and coverage greater than 70%.

Similarity Stage: Generating additional coding models using proteins from related species

UniProt alignments from the Raw Compute step were filtered to proteins classed by UniProt's Protein Existence (PE) classification as level 1 or 2. The proteins were also divided taxonomically into the following groups: mammalian, non-mammalian vertebrates and invertebrates. WU-BLAST was rerun for these sequences and the results were passed to Genewise to build coding models. Only models from the mammalian and non-mammalian vertebrate groups were used. The generation of transcript models using data from related species is referred to as the "Similarity stage". This stage resulted in 95,917 coding models.

Filtering Coding Models

Coding models from the Similarity stage were filtered using modules such as TranscriptConsensus. RNA-Seq spliced alignments supporting introns, models from the targeted stage, RNA-Seq models, pig cDNA and pig EST alignments were used to help filter the set. 47,326 models were rejected as a result of filtering, leaving 48,550 models used in subsequent stages. The Apollo software (Lewis et al 2002) was used to visualise the results of filtering.

RNA-Seq models

RNA-Seq data provided by the Swine Genome Sequencing Consortium (SGSC) was used in the annotation. This comprised a mixture of single and paired end data from samples including: a pool of 10 tissues, alveolar macrophages, male gonad, whole blood, placenta and testis. The available reads were aligned to the genome using BWA, resulting in 521,680,282 reads aligning. Subsequently, the Ensembl RNA-Seq pipeline was used to process the BWA alignments and create a further 7,800,861 split read alignments using Exonerate. The split reads and the processed BWA alignments were combined to produce

29,362 transcript models in total. The predicted open reading frames were compared to Uniprot Protein Existence (PE) classification level 1 and 2 proteins using WU-BLAST. Models with no BLAST alignment or poorly scoring BLAST alignments were split into a separate class.

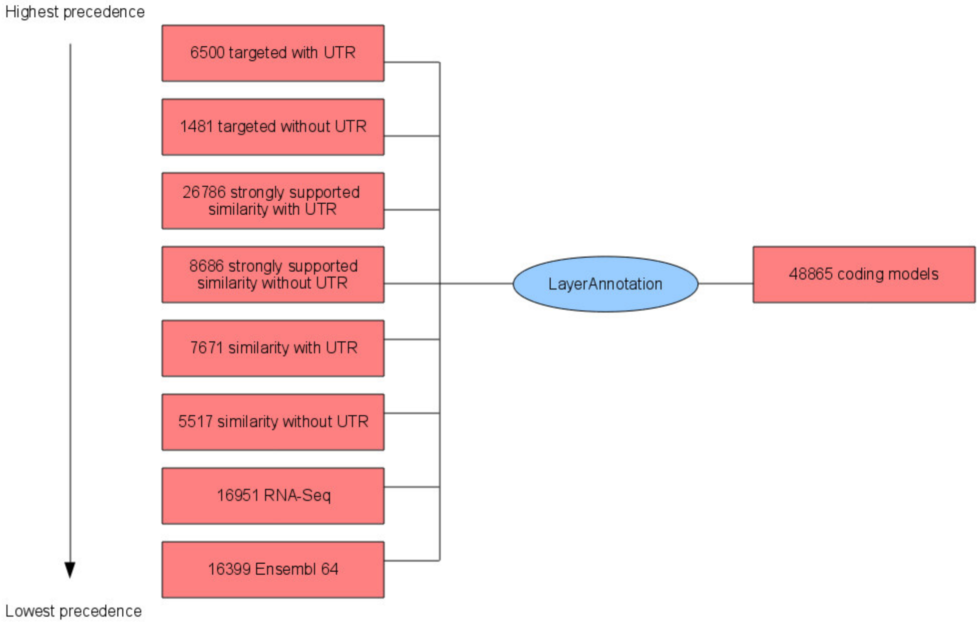
Untranslated region (UTR) addition

UTR were added to models generated in the targeted and similarity steps using the UTR addition modules and evidence from the pig cDNA models and all RNA-Seq models.

Layering of evidence

To combine models from different sources the LayerAnnotation module was used (Supplementary Fig. 5). This takes models from lower layers only where there are no models in a layer with higher priority. The layers, in order of precedence were: targeted with UTR, targeted without UTR, strongly supported similarity with UTR, strongly supported similarity without UTR, similarity with UTR, similarity without UTR, RNA-Seq with good BLAST scores to UniProt, and pig models from Ensembl 64, which were filtered to remove those based on projections in e64 and those with non-consensus splicing in e64.

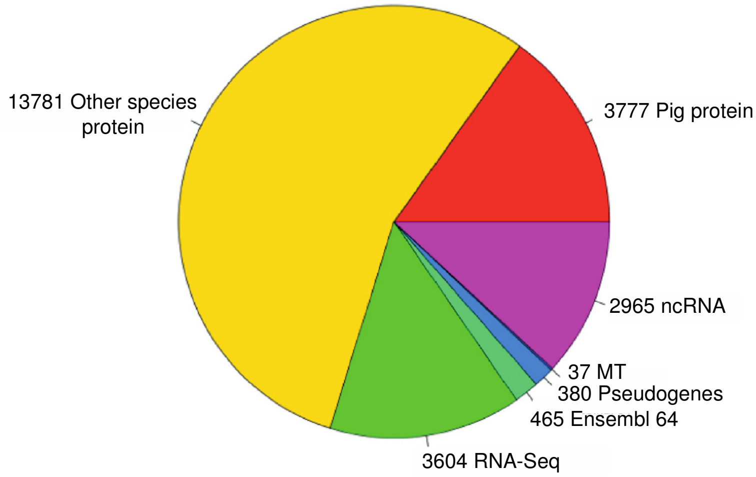
This led to a set of transcript models containing 7,655 from the targeted step, 36,119 from the similarity step, 3,630 from RNA-Seq evidence and 480 from the e64 ENSSSCP set



Supplementary Fig. 5: Combining evidence in LayerAnnotation

Generating multi-transcript genes

The above steps generated a large set of potential transcript models, many of which overlapped one another. Redundant transcript models were removed and the remaining unique set of transcript models were clustered into multi-transcript genes where each transcript in a gene has at least one coding exon that overlaps a coding exon from another transcript within the same gene. The resulting set of genes included 3,807 genes built using evidence from the targeted stage, a further 14,118 genes built using proteins from other species and 3,637 genes built only from RNA-Seq evidence.



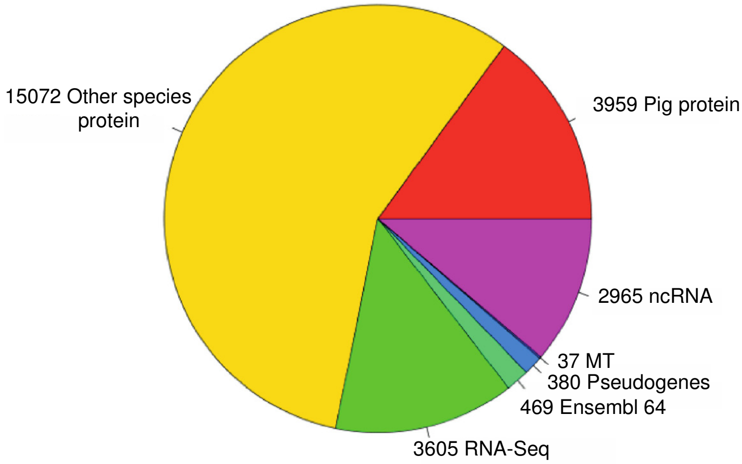
Supplementary Fig. 6: Composition of pig gene set.

Pseudogenes, protein annotation, non-coding genes, cross referencing, stable identifiers

The gene set was screened for potential pseudogenes. Before public release the transcripts and translations were given external references, or cross references, to external databases, while translations were searched for domains/signatures of interest and labeled where appropriate. Stable Identifiers were assigned to each gene, transcript, exon and translation. (When annotating a species for the first time, these identifiers are auto-generated. In all subsequent annotations the stable identifiers are propagated based on comparison of the new gene set to the previous gene set.)

Small structured non-coding genes were added using annotations taken from RFAM (Griffiths-Jones et al 2003) and miRBase (Griffiths-Jones et al 2006).

The final gene set consists of 21,640 protein coding genes (Supplementary Fig. 6), including mitochondrial genes, these contain 23,118 transcripts. A total of 380 pseudogenes were identified and 2,965 non-coding RNAs (ncRNAs). Of the protein coding transcripts 3,605 were made from RNASeq only, 15,072 came from proteins from other species. 3,959 transcripts were pig specific, 37 transcripts were mitochondrial (Supplementary Fig. 7).



Supplementary Fig. 7: Composition of pig transcripts

Further information

The Ensembl gene set is generated automatically (Curwen et al.2004; Potter et al. 2004), meaning that gene models are annotated using the Ensembl gene annotation pipeline. The main focus of this pipeline is to generate a conservative set of protein-coding gene models, although non-coding genes and pseudogenes may also be annotated.

Every gene model produced by the Ensembl gene annotation pipeline is supported by biological sequence evidence (see the "Supporting evidence" link on the left-hand menu of a Gene page or Transcript page); *ab initio* models are not included in our gene set. *Ab initio* predictions and the full set of cDNA and EST alignments to the genome are available on our website.

More information on the Ensembl automatic gene annotation process can be found at:
http://www.ensembl.org/info/docs/genebuild/genome_annotation.html
http://cvs.sanger.ac.uk/cgi-bin/viewvc.cgi/ensembldoc/
pipeline_docs/the_genebuild_process.txt?root=ensembl&view=co

3. Non coding RNAs

Gene content

Using a computational pipeline to search for sequence and structure homology (Anthon C, Tafer H, Havgaard JH, Seeman S, Stadler P Nielsen M, Nilsen RO, Hedegaard J, Fredholm M, Thomsen B and Gorodkin J , personal communication), we identified a total of 139 cis regulatory mRNA elements and 3,276 ncRNAs including 374 microRNAs (miRNAs), 185 ribosomal RNAs (rRNAs), 1,030 small nuclear RNAs (snRNAs), 638 small nucleolar RNAs (snoRNAs), 819 transfer RNAs (tRNAs) and 219 ncRNA genes belonging to other families (Gardner et al. 2010; Kozomara et al. 2011; Juhlimng et al. 2009; Lestrade et al. 2006; Pruesse et al. 2007). In addition, we identified 186 novel miRNA candidate genes in a small-RNA dataset. We predicted 459 miRNA loci with miRDeep (Friedlander et al. 2008) based on RNAseq data, of which 273 overlapped with ncRNA candidates predicted by the pipeline for sequence and structure homology. This resulted in an additional 186 miRNA candidate loci, yielding a total of 560 miRNAs in 490 corresponding families in the porcine genome. Among the miRDeep predicted miRNAs 254 showed signals for of both mature and star sequences.

Curation

Certain RNAs (7SK RNaseP_nuc SNORD13/snoU13 U1 U11 U12 U2 U3 U4 U5 U6 U6atac U7 U8) are known to be polymerase II (pol-II) or polymerase III (pol-III) transcripts and these RNAs were curated through identification of active pol-II and pol-III transcripts which for the snRNAs led to a reduction from 1030 to 70 loci. The snoRNAs were curated by comparing the porcine host genes with that of the human homolog leading to the confirmation of 159 loci (155 snoRNAs and 4 conflicts). The curation of cis-regulatory mRNA elements led to 85 being close to or overlapping with protein coding genes on the same strand. As a part of our annotation and curation efforts we labelled each ncRNA with its protein context and with the conservation of the locus in 20 vertebrate species.

ncRNA summary

A summary of the ncRNA annotation obtained by the homology search pipeline and the subsequent curation effort is presented in Supplementary Table 4. The first column is the classification of the RNAs: “cisreg” are cis-regulatory elements from Rfam. “miRNA” are miRNAs obtained by sequence homology from mirBase. “rRNA” are ribosomal RNAs mainly found with RNAmmer. “snRNA” and “snoRNA” are merged results for small nuclear RNAs and small nucleolar RNAs from sequence and structure homology search. “tRNA” are merged results, but are mainly found with tRNA-scan. “other” are non-coding RNAs from Rfam that do not belong to one of the classes above. “conflicts” are conflicts of annotation. The second column contains the number of genes curated by methods explained in the text. The third column contains the number of the pseudo-genes detected by identification of pol-II and pol-III transcripts. The fourth column contains the number of the miRNAs discovered by miDeep. The first number is the new candidates and the second is the overlap with the annotation pipeline. The fifth column is the genes from the annotation pipeline and from mirDeep that remained after the curation procedure. The full ncRNA annotation is available for download from http://rth.dk/resources/rnannotator/susscr102.

Supplementary Table 4. Summary of the ncRNA annotation

	Curated genes	Pseudo-genes	mirDeep detected	Curated annotation	High confident annotation
cisreg	85	-	-	85	139
miRNA	-	-	186 (264)	560	560
rRNA	-	-	0 (1)	185	185
snRNA	69	960	-	70	1030
snoRNA	155	-	0 (6)	155	638
tRNA	-	-	-	819	819
other	5	128	-	91	219
conflicts	4	-	0 (2)	11	11
Sum	318	1089	186 (273)	1976	3601

4. Manual annotation of the genomic complement of porcine immune genes

Our comprehensive and detailed annotation of the porcine genome assembly reveals evidence for specific gene family expansions, gene duplications, and high levels of positive selection in porcine genes related to immunity.

Immune genes are known to be actively evolving in the human and other species (Barreiro and Quintana-Murci, 2010; Bovine Genome Sequencing and Analysis Consortium 2009; hereafter BGSAC, 2009), and the new porcine assembly was used to investigate evidence for gene family expansion, gene duplication, and positive selection of protein sequences.

Extensive manual annotation of the genomic complement of porcine immune genes

The Immune Response Annotation Group (IRAG) members have used Otterlace (Searle et al. 2004; Loveland et al. 2012) to manually annotate over 1,400 loci in build 9 (Sscrofa9) selected based on their membership in immune pathways or Gene Ontology immune response annotation. These annotations will be automatically ported to Sscrofa10.2 and discrepancies addressed in 2012 (Supplementary Table 5). Members confirmed automated annotation of 1,022 known genes through manual annotation of 3,873 transcripts and 1,469 gene models; 1,738 annotated transcripts contained the full protein coding sequence. Importantly, the cross-species alignment tools in Otterlace allowed annotation of 1,218 transcripts in the pig genome using only RNA data from other species (Supplementary Table 5). Fifty pseudogenes were also identified during the annotation of these genes, 20 of which map to swine chromosome 7 (SSC7).

Examination of the swine major histocompatibility complex (MHC) or swine leukocyte antigen (SLA) genes on SSC7 revealed new features for the SLA class I genes. For the SLA complex, a reference haplotype Hp1a.1 spanning the entire region has been fully sequenced and annotated (Renard et al., 2006) and is available at http://vega.sanger.ac.uk/Sus_scrofa/. Two additional reference class I haplotypes, Hp28.0 and Hp62.0, have been characterized affirming the existence of another SLA-Ia gene, SLA-12 (Tanaka-Matsuda et al., 2009). From these data, four functional classical class I genes (SLA-Ia) SLA-1, -2, -3 and -12 have been identified together with four pseudogenes, SLA-4, -5, -9 and -11, showing that SLA haplotypes differ by copy number variations of the SLA-Ia genes and pseudogenes. In the current assembly, the SLA-5 gene was found, verifying that this gene can be coding in some haplotypes, as suggested (Renard et al., 2006; Tanaka-Matsuda et al., 2009). Thus, in pigs SLA-5 has to be included as new SLA-Ia gene. The three expressed non-classical class I genes (SLA-Ib), SLA-6, SLA-7 and SLA-8, have unknown functions but are considered as putative functional homologs of the human genes known as HLA-E, -F and -G (Renard et al., 2006, Kusza et al., 2011). It is clear from the genome build and from released transcripts that SLA-11 is not a pseudogene and has alternative transcripts that look like those found for SLA-6, meaning that SLA-11 is a new, not yet, identified SLA-Ib gene. The presence of SLA-11 in the SLA-Ia gene cluster affirms that SLA-Ia and -Ib genes are intermingled in the pig, as in humans, contrary to previous reports (Renard et al., 2006). Due to the complexity of the SLA region, and since the sequenced genome was not homozygous at the SLA locus, the 10.2 genome build cannot serve as a new reference SLA haplotype but is useful to refine the gene content of the SLA complex. It is anticipated that, as in humans, targeted resequencing of numerous SLA haplotypes will be required to provide accurate haplotype-specific information (Horton et al., 2008). A striking feature of the SLA complex is the physical separation of the region by the centromere that is unique among mammals studied to date. This unique feature provided an opportunity to address the impact of centromeric sequences on gene expression; indeed, decreased transcriptional activity was observed in the centromeric regions (Gao et al., manuscript in preparation).

Supplementary Table 5. Summary of Genome Annotation of IRAG Project: Use of Sequences from Other Species adds over 1,200 transcripts (>30%)

Chromo some	Genes*	Known genes annotated	Tanscripts annotated	Protein coding transcripts annotated	Complete protein coding transcripts annotated	Predicted genes found to be pseudogenes**	Non-organism-supported*** transcripts annotated
1	121	82	228	196	130	9	66
2	110	84	302	239	131	1	90
3	79	53	176	137	68	0	36
4	125	93	364	298	202	7	112
5	62	54	173	138	88	0	51
6	91	66	226	176	100	1	84
7	232	172	603	360	242	20	58
8	47	34	114	93	38	0	24
9	66	45	135	107	65	2	41
10	21	19	84	72	30	0	17
11	12	10	18	16	9	0	4
12	70	58	167	138	63	3	63
13	62	47	170	136	64	1	68
14	89	61	285	222	129	2	117
15	38	30	110	85	40	1	45
16	24	16	58	45	18	0	28
17	54	38	105	88	48	3	22
18	14	10	20	18	9	0	5
X	152	50	535	445	264	0	287
Total	1469	1022	3873	3009	1738	50	1218

* Number of gene objects created in the Otterlace annotation system

** Processed and non-processed pseudogenes are included

*** No porcine EST or cDNA sequence was available to create these transcript predictions

T cell receptor (TCR) genes possess sequences with high repetitions, so that it is difficult for shotgun sequencing with low redundancy or sequencing with short-read next-generation sequencers to generate correct sequences of such loci. Therefore, intensive efforts on sequencing of the TRA/TRD (T cell receptor α/δ) and TRB (T cell receptor β) loci have been conducted in pigs. The pig TRD locus is embedded in TRA, and D (diversity) (Dδ) and J (joining) segments (Jδ), and genes encoding C (constant) region of TCR δ (Cδ) are located between the V (variable) segments of TCR α/δ (Vα/Vδ) and J genes of TCR α (Jα), as observed in other mammals. All of the human 61 Jα segments correspond to those of pig, and most of mouse Jα can be allocated to orthologs in pig. These indicate functional similarity of the TCR α molecule between human, mouse and pig (Uenishi et al. 2003). On the other hand, the pig TCR δ gene (TRD) has a more complicated structure than those of human and mouse. Pig has at least 6 Dδ genes, while human and mouse have 3 and 2, respectively. The pig Dδ genes are frequently used in functional TCR δ transcripts with up to 4 concatenated domains (Uenishi et al. 2009). This indicates that the pig can generate a high diversity of TCR δ chain molecules to cope with antigens, which may be related to the fact that the percentage of γδ T cells in peripheral blood is much higher in pig than that of human and mouse (Binns et al. 1992). As for TRB, pig has 3 functional Dβ-Jβ-Cβ units, while human and mouse have 2 units, respectively (Eguch-Ogawa et al. 2009).

The immunoglobulin heavy chain (IGH) locus organization on SSC7 was described previously (Eguchi-Ogawa et al 2010). IGHV gene diversity is highly restricted, as in cattle, but all known porcine IGHV genes belong to a single family, IGHV3, whereas cattle IGHV are restricted to IGHV4 (Saini et al 1997). The lambda light chain (IGL) locus on SSC14 contains 22 IGLV gene segments, with 9 appearing functional. The locus is organized into two distinct clusters, a constant (C)-proximal cluster containing IGLV3 family members, and a C-distal cluster containing IGLV8 and IGLV5 family members (Schwartz et al 2012a). The porcine IGLV8 subgroup genes have recently expanded, suggesting a particularly effective role in immunity to porcine-specific pathogens, especially since IGLV expression is nearly exclusively restricted to the IGLV3 and IGLV8 (Butler et al 2006, Schwartz et al 2012a). The locus also contains three non-functional IGLV1 that are orthologous to cattle IGLV, all of which are in this family (Pasman et al 2010).

The IGL locus contains three tandem cassettes of IGLJ and IGLC, two of which are functional, and a fourth IGLJ with no corresponding IGLC. The kappa light chain (IGK) locus on SSC3 is comprised of a single IGKC gene and five IGKJ genes that lie 27.9 kb downstream from the IGKV genes (Schwartz et al 2012b). This locus contains at least 14 IGKV genes, of which 9 are functional and belong to either the IGKV1 or IGKV2 gene families. Polymorphisms within the individual Duroc sow, that was sequenced, revealed alleles that differed by as much as eight percent in amino acid sequence among IGLV genes and by as much as 16 percent among IGKV genes, indicating that allelic variation may substantially expand the diversity of the porcine antibody repertoire (Schwartz et al 2012a, 2012b). Many IGKV2 CDR1 are shared between genes but not between alleles. Thus, germline gene conversion may help develop a high level of IGK allelic variation.

Immune Gene Family Expansion

We targeted several gene families for a detailed analysis of expansions across species; families were chosen from a preliminary analysis done in humans, mice and pigs (Dawson 2011). Artiodactyl-associated families were included based upon expansions noted in the bovine genome (BGSAC, 2009). Unambiguous 1:1 orthologs for each species were initially determined from the corresponding human or mouse gene in Ensembl. For each gene, additional family members were determined by including genes that were listed as ambiguous orthologs (1:many) or by a separate Ensembl within species search for paralogs. Each Ensembl predicted gene transcript was BLASTed against the NCBI reference sequence database to determine the corresponding NCBI loci and reference sequence or other family members that may have been missed due to areas of the genome that were not sequenced. Results for the cathelecidins are shown in Supplementary Table 6 .

Supplementary Table 6. Comparison of Immune Gene Family membership across four mammals The comparison shows evidence of Porcine-specific and Cetartiodactyl-specific Expansion.

Family Description	**Number of genes found for each family per species***			
	Human	**Mouse**	**Cow**	**Pig**
Proposed Cetartiodactyl-specific Expansions				
Cathelicidin Superfamily	1	1	10	10
Type I Interferon (inclusive)	17(12)	25(2)	51(13)	39(16)
Type I Interferon, Alpha Subfamily	13(4)	13(1)	19(3)	18(9)
Type I Interferon, Omega Subfamily	1(8)	0	19(7)	7(5)
Proposed Ruminant -specific Expansions				
Beta Defensin Superfamily	39(9)	51(1)	~106(7)	34(2)
C-type Lysozyme/LYZ1 Superfamily	9	9	16	7
Type I Interferon, Beta Subfamily	1	1	8(1)	1
Type I Interferon, Tau Subfamily	0	0	4(2)	0
Proposed Porcine -specific Expansion				
Type I Interferon, Delta Subfamily	0	0	0	11(2)
Additional Immune Gene Families Annotated				
BPI Superfamily	12(2)	16	18	14(2)
CCL Chemokine	28(1)	39(5)	22	21
CD1 Superfamily	5	2	15(2)	4(1)
CD163/WC1 Superfamily	3	4	15	4
CLECT Superfamily (inclusive)	50(4)	90(6)	55	44
CLECT Superfamily, Asialoglycoprotein and DC Receptor Subfamily	16	24(1)	13	13
CLECT Superfamily, Collectin Subfamily	7(2)	7	10	7
CLECT Superfamily, NK Cell Receptor Subfamily	22(1)	52(5)	29	21
CLECT Superfamily, Reg Subfamily	5	7	3	3
GH18 Chitinase Like Superfamily	6(1)	9(1)	8	7
NLR and Pyrin Superfamily	30(4)	43(8)	23	24
Resistin Superfamily	2	4	1	2
RNase A Family	14	22(6)	16(1)	13(1)
S100 Superfamily	21	17(1)	19	20
SAA Superfamily	4(1)	5	6	6
Toll Like Receptor	10(3)	12(1)	10	10(2)
TRIM E3 Ubiquitin-protein Ligase Superfamily, TRIM5 Subfamily	4	10(1)	5	3
Type I Interferon, Epsilon Subfamily	1	1	0	1
Type I Interferon, Kappa Subfamily	1	1	1	1
Type I Interferon, Zeta Subfamily	0	9(1)	0	0
ULBP Superfamily	6	2	12	7

*Numbers of confirmed pseudogenes are shown in parentheses

We have identified a number of expansions in porcine gene families important in the immune system adding important evolutionary information on the immune response. For the bovine genome the Bovine Genome Sequencing and Analysis Consortium (BGSAC, 2009). documented expansions of the cathlecidin, beta-defensin, interferon, and C-type lysozyme gene families, among others. Our porcine genome analyses show that some of these expansions are present in the porcine genome and thus provide evidence for an artiodactyl-specific expansion, as well as evidence that others of these expansions are not present in the pig genome, providing additional data for a ruminant-specific expansion as proposed (BGSAC, 2009).

(a) Both cattle and pigs have 10 cathelicidins compared to only one in human and mouse. For cathlecidins, then, we document a potentially artiodactyl-specific expansion relative to human and mouse genomes, not a ruminant-specific expansion as proposed (BGSAC, 2009)

(b) On the other hand, while cattle have many (>100) beta-defensin genes, only 34 beta defensin genes have been detected in the current swine genome assembly. This number is comparable to that seen in the human genome (39). A similar result is observed for the C-type lysozyme family in pigs, which has 7 genes, while 16, 9 and 9 are found in the bovine, human and mouse genomes, respectively. Thus, our analysis of the second artiodactyl genome indicates that beta-defensin and C-type lysozyme family expansions observed in cattle may be ruminant-specific adaptations (BGSAC, 2009).

(c) An interesting discovery is that pigs have at least 39 type I interferon (IFN) genes, which is double the number in human and significantly more than seen in mouse. We also found 16 pseudogenes in this family. Cattle have 51 type I IFNs (13 pseudogenes), indicating that both bovine and porcine type I IFN loci are actively undergoing duplication.

Gene Duplication

In the course of annotating the immune response gene families shown in Supplementary Table 7, we have found evidence of gene duplication and pseudogenes in the build 10.2 assembly. A summary is shown in Supplementary Table 6, with over 150 gene sequences analyzed. Using extreme sequence similarity (99-100% identity) as a metric, we propose that many of these duplications (>125) are due to assembly artifacts. However, we have identified clear evidence for duplication of six immune-related genes: *IL1B*, *CD68*, *CD163*, *CRP*, *OAS1*, and *IFIT1*, and one non-immune gene, *RDH16*.

Supplementary Table 7 Gene duplications in immune response gene families. *Genes RDH16 and RDH16L1 represent non-immune related genes.

Gene	NCBI Locus	Status	Ensembl Build 9 Gene ID	Chr	Ensembl Build 9.2 Gene Coordinates	Ensembl Build 10.2 Gene Coordinates	Comments
CD36	733702	Biological Duplication	ENSSSCG00000015405	9	93,201,790-93,249,833	NC_010451.3 :110040077-110102645	
CD36L1	100511343		NA	9	NA	NC_010451.3 :109954298-110015639	Sequences in Unigene Ssc.35355 do not align to an identified gene in Ensembl build 9.2
CRP	396842	Biological Duplication	ENSSSCG00000006403	4	94,315,127-94,322,791	NC_010446.4 :98764964-98772727	Gene is not in build 9.2 but is in build 10.2f.
CRPL1	100620468		NA		NA	NC_010446.4 :98745242-98757413	Porcine-specific gene.
IL1B	397122	Biological Duplication	ENSSSCG00000008088	3	38,955,442-38,962,433	NC_010445.3 :45319296-45326287	
IL1BL1	396565		ENSSSCT00000008860	3	38,885,708-38,886,829	NC_010445.3 :45179378-45185849	
CD68	100522571	Biological Duplication	ENSSSCG00000017956	12	50,060,706-50,063,245	NC_010454.3 :55296986-55299419	Gene on porcine chromosome 12 is syntenic with human CD68 on chromosome 17.
CD68L1	100520753		ENSSSCG00000008769	8	23,617,243-23,619,485	NC_010450.3 :29672313-29674595	Gene on porcine chromosome 8 is syntenic with human chromosome 4.
CD163	397031	Biological Duplication	NA	Un		NW_003539177.1 :5-471	Sequences in Unigene Ssc.5053 do not align in Ensembl build 9.
CD163L1	100144477		NA	5		NC_010447.4 :65984979-66021167	CD163L1 is duplicated in cows, humans and mice.
CD163L2	100627089		NA	5	NA	NC_010447.4 :66030456-66085371	CD163L2
IFIT1	100153038	Biological Duplication	ENSSSCG00000010453	14		NC_010456.4 :110223576-110235661	
IFIT1L1	100621926		NA	1		NC_010443.4 :296319106-296331356	
RDH16*	100511633	Biological Duplication	ENSSSCG00000000414	5		NC_010447.4 (24133013..24139006	Artiodactyl-and Perissidactyl-specfic duplication, there are 3 putative bovine and 3 equine genes that are similar to RDH16.
RDH16L1*	100512656		NA	5		NC_010447.4 (24087398..24094529	Sequences in Unigene Ssc.94004 or Ssc.55153 do not align to an identified gene in Ensembl build 9.

Accelerated selection of immune related genes

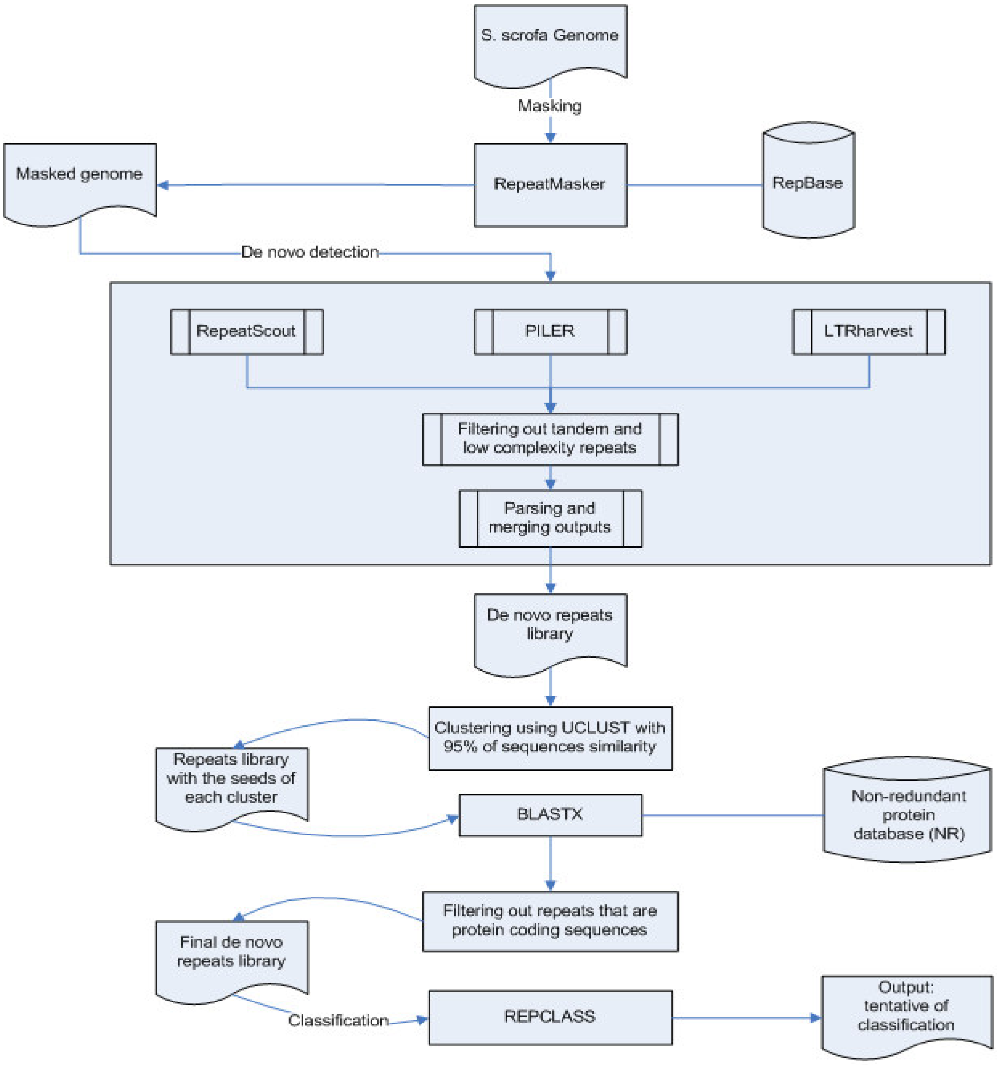
A randomly selected subset of 158 immunity-related pig proteins from the IRAG annotated gene set was studied for evidence of positive selection using the PhyleasProg phylogenetic analysis web server (http://phyleasprog.inra.fr/; Busset et al., 2011). Based on Ensembl release 64, it enables users to reconstruct phylogenetic trees, calculate accelerated evolution with a visualization of these results on the protein sequence and on a 3D structure where possible, and explore genomic environment of query genes. To calculate accelerated evolution we used site models and branch-site models of PAML (Yang et al., 2007). Branch-site models are designed to detect signals of local episodic positive selection in order to determine whether different species underwent selective pressure. In new version 2.6, multiple sequence alignments can be performed by MUSCLE (Edgar, 2004) or by PRANK (http://www.ebi.ac.uk/goldman-srv/prank/), and are refined by GBLOCKS (Talavera and Castresana, 2007), as improved by a local Perl script.

Supplementary Table 8 see excel file "SupTable8.xls"

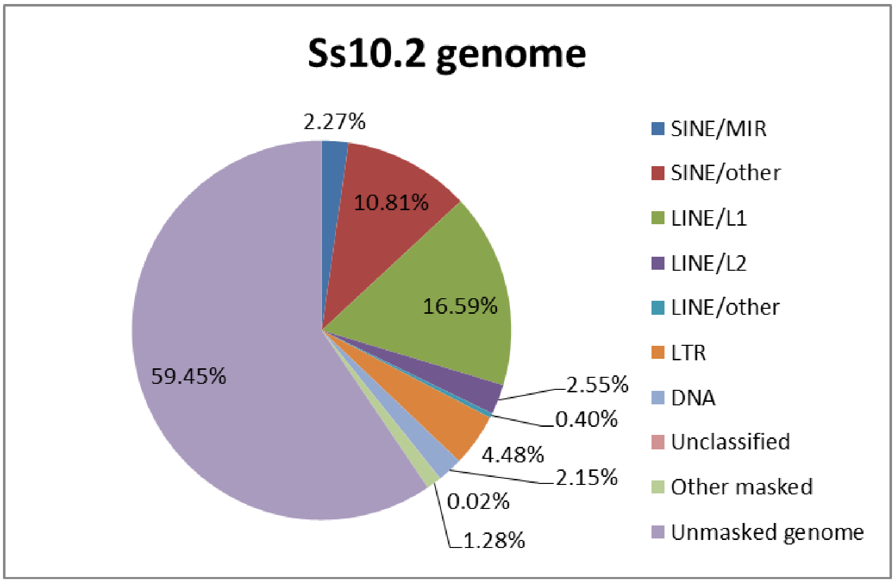
5. Repetitive elements and PERVs

Discovery of novel repetitive elements

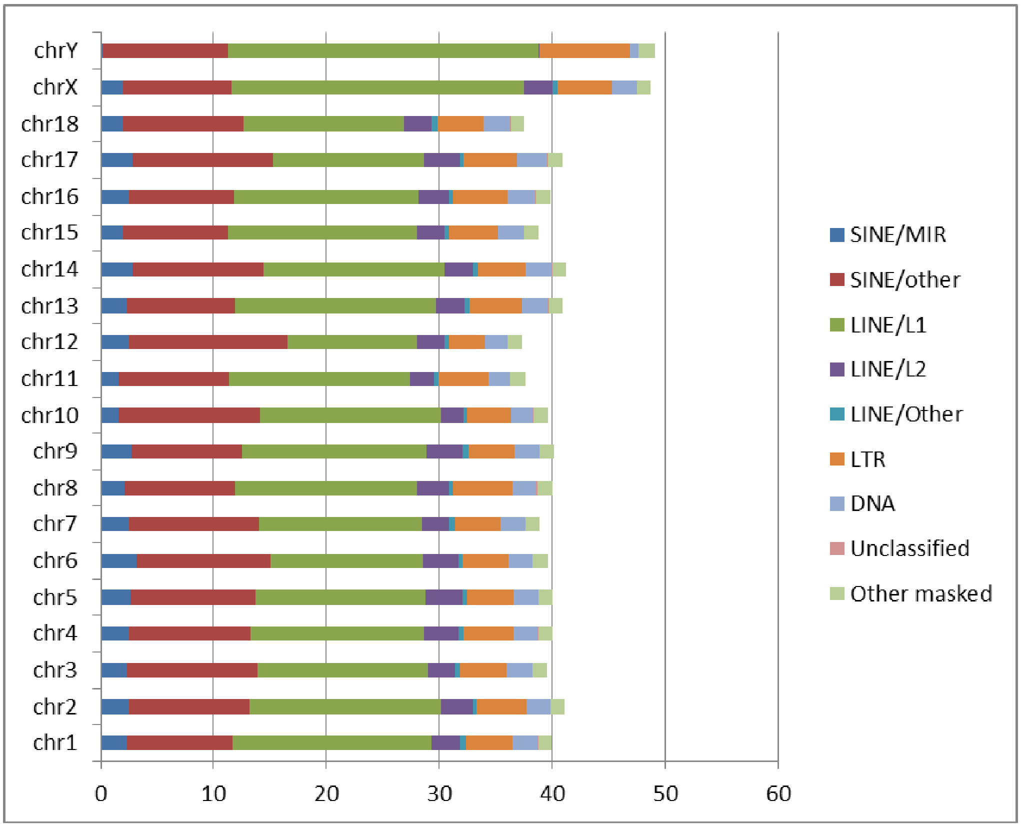
A *de novo* repeat identification strategy was applied to the Sscrofa10.2 assembly. Briefly, the genome was first masked with RepeatMasker (Smit et al; Repeatmasker) to remove regions with high similarity to previously known repeat families. PILER (Edgar & Myers, 2005), RepeatScout (Price et al 2005) and LTRharvest (Ellinghaus et al 2008) were then applied to the masked assembly. Tandem repeats and low complexity regions were removed, followed by clustering, to produce a merged list of putative *de novo* repeat families. These were then clustered by UCLUST (UCLUST) at 95% identity thresholds; consensus regions were then compared with TBLASTX against the non-redundant (NR) NCBI protein sequence database to remove sequences matching known peptides. Chromosome-specific repeats were removed as likely segmental duplications. The remaining sequences were then classified by REPCLASS (Feschotte et al 2009), with the annotations manually curated and uploaded to RepBase (Repbase; Jurka et al 2005). The *de novo* repeat discovery and annotation strategy is summarised in Supplementary Fig. 8. RepeatMasker was then used to re-mask Sscrofa10.2 using the updated RepBase reference. The results are summarised in Supplementary Figs 9-11.


S. scrofa Genome
Masking
RepeatMasker
RepBase
Masked genome
De novo detection
RepeatScout
PILER
LTRharvest
Filtering out tandem and low complexity repeats
Parsing and merging outputs
De novo repeats library
Clustering using UCLUST with 95% of sequences similarity
Repeats library with the seeds of each cluster
BLASTX
Non-redundant protein database (NR)
Filtering out repeats that are protein coding sequences
Final de novo repeats library
Classification
REPCLASS
Output: tentative of classification

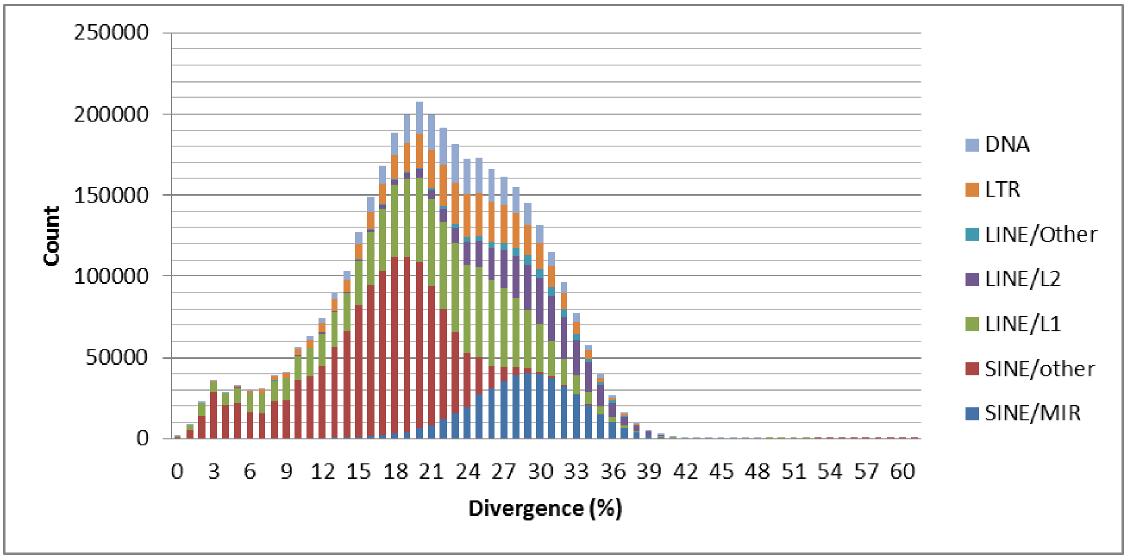
Supplementary Fig. 8 Summary of *de novo* repeat discovery and annotation strategy



Supplementary Fig. 9 Repetitive sequence composition of the porcine genome



Supplementary Fig. 10 Repetitive sequence composition of the individual chromosomes of the porcine genome

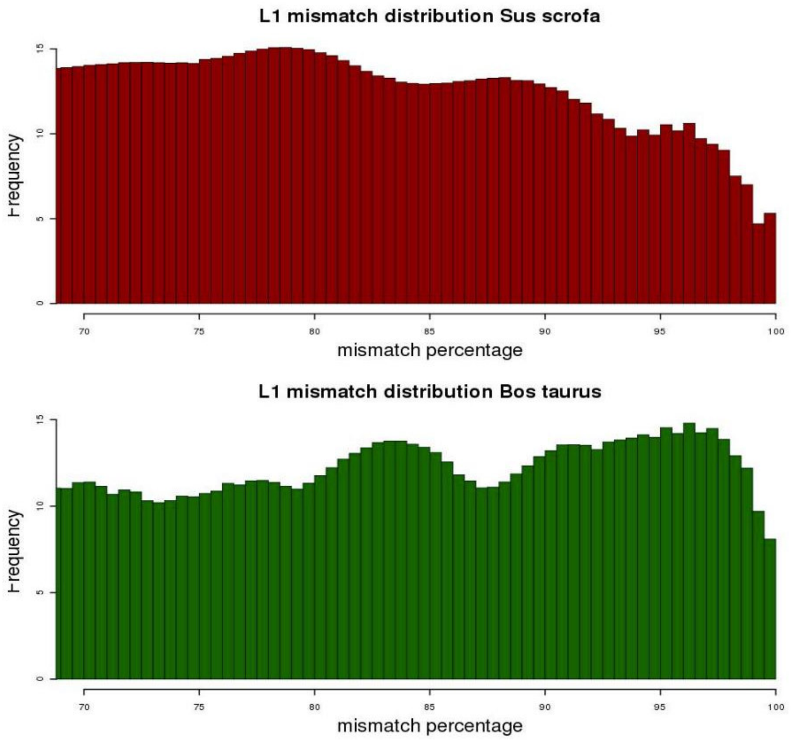


Supplementary Fig. 11 Sequence divergence of porcine **r**epetitive sequence families

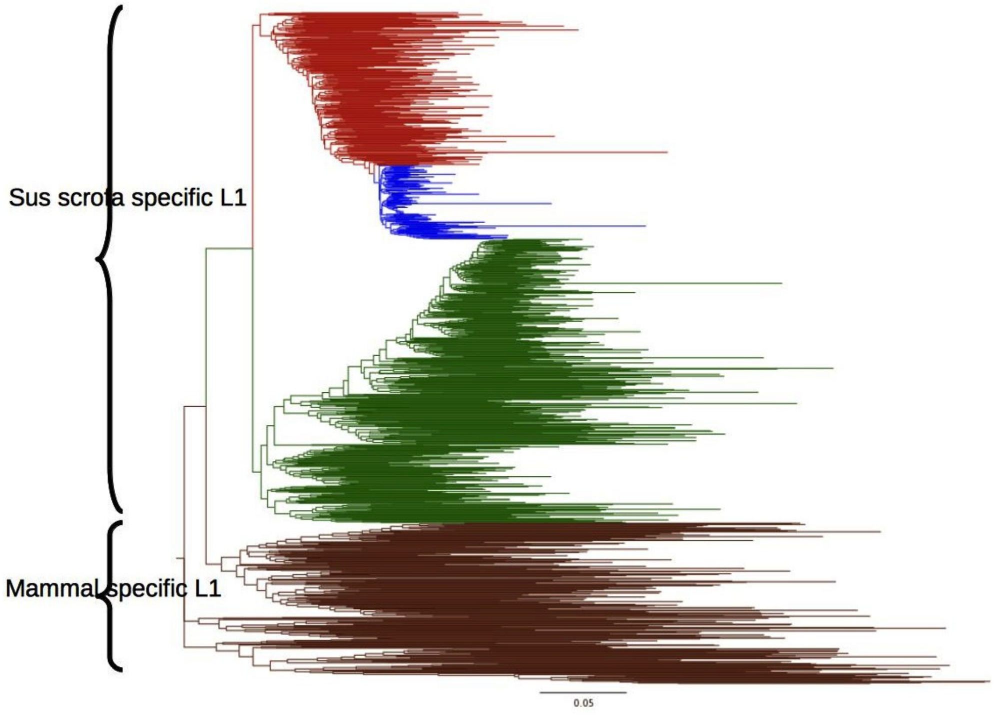
From supplementary Figure 11 it is clear that long-interspersed-repeat 1 (LINE1) and glutamic acid transfer RNA (tRNA$^{Glu)}$-derived small-interspersed-repeats (SINEs) also referred to as porcine repetitive element (PRE) (the majority of those classified as SINE/Other) are not only the groups that are most abundant, but are also showing signs of an ancient expansion and recent inactivation, as can be deduced from the sequence divergence.

This pattern is contrary to what is found in other artiodactyls such as cattle, that have experienced recent high activity of repetitive elements such as LINE1, as can be seen in Figure 12, where pairwise mismatch percentages of 1,000 randomly sampled annotated LINE1 elements were compared. Although we did not formally date the initial expansion of LINE1 elements, the mismatch percentages of ~20% near the largest peak suggests a post-Cretaceous expansion.

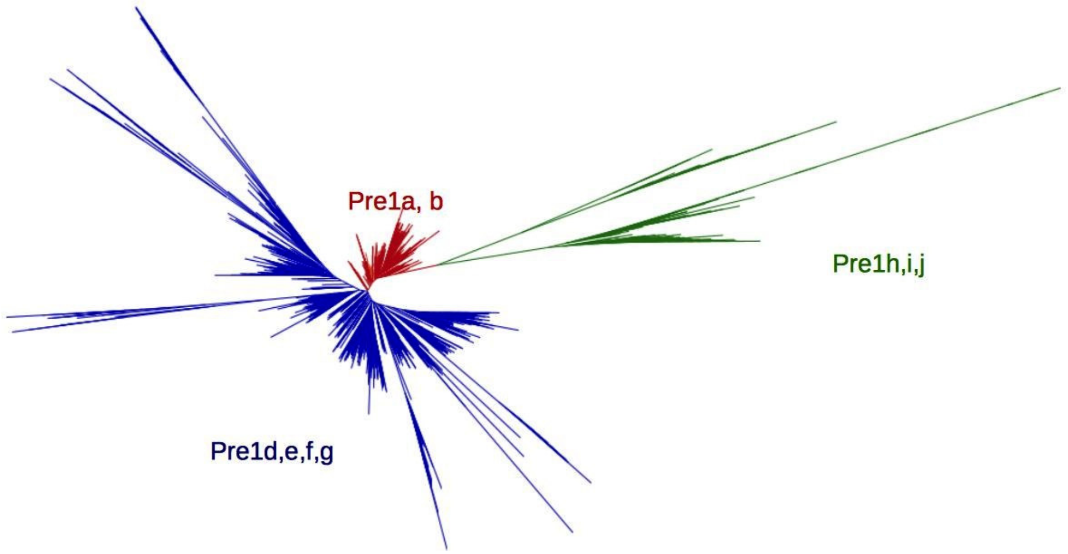
A phylogenetic analysis of 1,000 randomly selected annotated elements of both LINE1 and PRE is shown in Supplementary Figures 12 and 13. All the LINE1 and PRE-1a elements that have a 3% or smaller pairwise divergence cluster together, suggesting that each currently is represented by a single active lineage



Supplementary Fig. 12: Comparison of L1 mismatch distribution shows that recent L1 activity has been far higher in the cow genome compared to the pig genome. The frequencies have been transformed by taking the square root to correct for over representing ancient elements.



Supplementary Fig. 13: Phylogenetic analysis of 1000 randomly sampled Porcine L1; only the blue lineage contains elements that are highly similar (<3% similarity) indicating recent and possibly current activity.

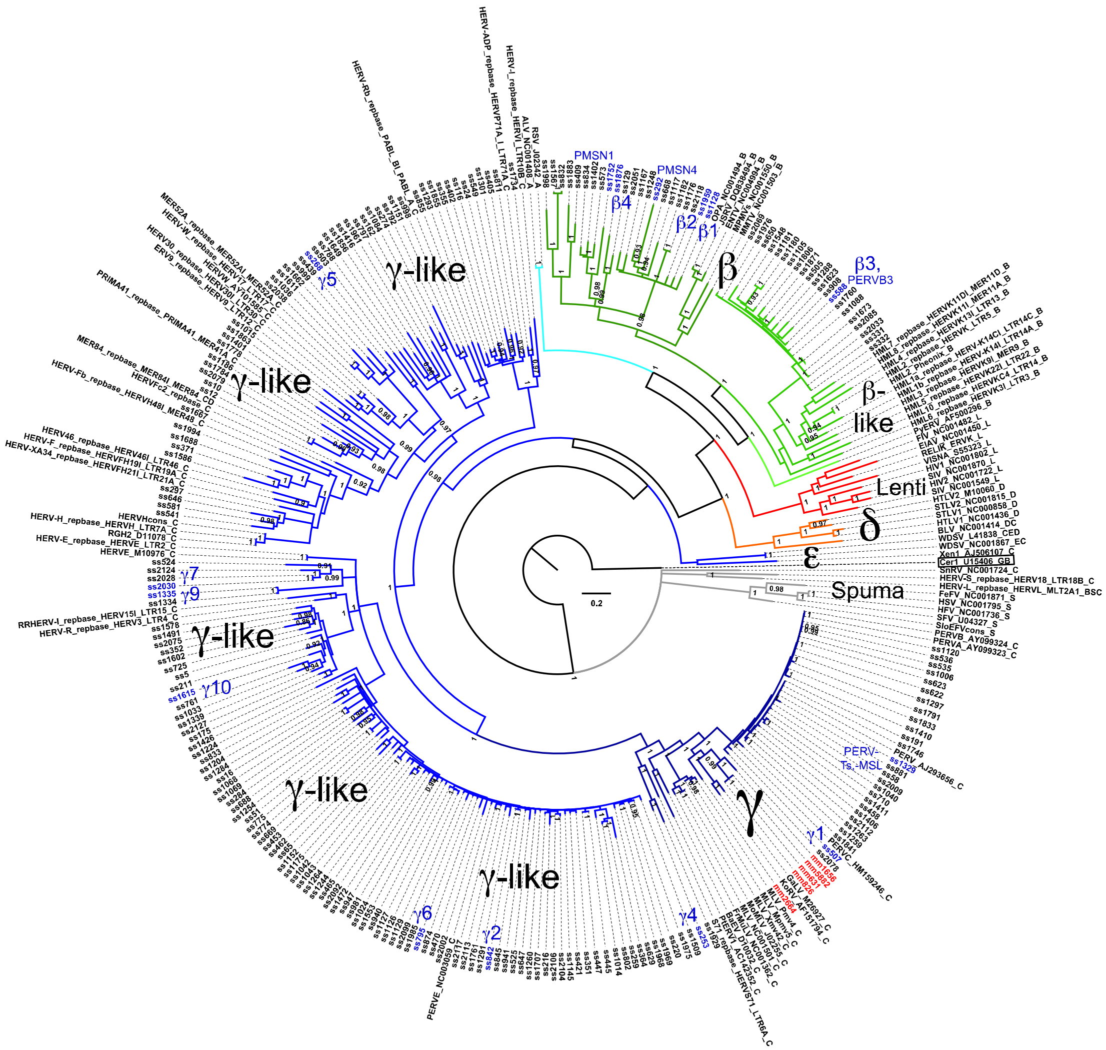


Supplementary Fig. 14: Phylogenetic analysis of 1000 randomly sampled Porcine PRE sequences. Only the PRE1a lineage contains elements that are <3% similar, indicating very recent and possibly current activity.

Porcine endogenous retroviruses (ERVs)

Porcine endogenous retroviruses (PERVs) are of considerable medical significance because they may give rise to viruses which can infect humans and thus pose a potential risk of zoonosis in pig-to-human xenotransplantation. We used the ERV specific RetroTector software (Sperber et al Nucl Acids Res 2007) and identified an additional 175 provirus insertions compared with Repeatmasker, highlighting the importance of using different methodologies for retroposon annotation (ssERVs, Supplementary Table 9).

We constructed a Bayesian phylogeny of 212 *Sus scrofa 10.2* ERVs representing 551 detected ERVs, in relation to 82 reference sequences (Genbank and Repbase) rooted on Cer1 (boxed). BLAST searches of ssERVs against all available vertebrate genome assemblies revealed no close similarity except between gamma 1 PERVs (PERV-A, -B, and -C) and a subset of mouse ERVs (*Mus musculus 8*), indicated in red in Supplementary Fig. 15. The majority of ssERVs appeared to be gamma-like with beta-like ssERVs being the second most represented group. Reference sequences provide a framework supporting current retrovirus group designations except for two discrepant beta-like clades.



Supplementary Fig. 15. Bayesian phylogeny of 212 *Sus scrofa 10.2* ERVs (ssERVs, Supplementary Table 9) representing 551 detected ERVs, in relation to 82 reference sequences (Genbank and Repbase) rooted on Cer1 (boxed). Closest matches to previously reported ssERV groups are indicated in blue while mouse ERVs are shown in red. (Beta1: AF274710, Beta2: AF274711, Beta3: AF274712, Beta4: AF274713, Gamma1: AF274705, Gamma2: AF274706, Gamma4: AF274708, Gamma5: AF274709, Gamma6: AF511106, Gamma7: AF511111, Gamma9: AF511113, Gamma10: AF511114, PERVB3: AF274712, PERVMSL: AF038600, PERVTs: AF038599, PMSN1: AF277320, PMSN4: AF277322). The phylogeny was constructed using the GTR+G model in MrBayes3.2 and run for 20 million iterations.

Supplementary Table 9 Positions (*Sus scrofa* 10.2) for ssERVs present in pylogeny

susScr_10. 2_ERV_id	Chr	Start	End	susScr_10. 2_ERV_id	Chr	Start	End
ss5	1	103486379	103491895	ss1151	9	47099908	47096612
ss10	1	108933858	108923873	ss1152	9	47129756	47138452
ss12	1	109156749	109164904	ss1167	9	57331402	57322540
ss16	1	111361920	111356036	ss1175	9	68296888	68305715
ss24	1	123664656	123655684	ss1176	9	68371186	68374784
ss51	1	142518716	142522635	ss1180	9	68527445	68523545

ss58	1	147287602	147295434
ss65	1	158383767	158375042
ss116	1	198466766	198459727
ss129	1	217997851	218004838
ss162	1	251693902	251701342
ss175	1	266519739	266505936
ss191	1	294481004	294488765
ss211	1	310100338	310112109
ss216	1	311928360	311920499
ss220	1	313237399	313230065
ss253	1	57495940	57506269
ss259	1	68594320	68591241
ss268	1	78313201	78302872
ss274	1	82996030	82990378
ss284	1	91018057	91009255
ss292	2	10441212	10435806
ss297	2	108873715	108869452
ss331	2	152130771	152122586
ss332	2	152256085	152247902
ss351	2	160332284	160325188
ss352	2	160538630	160548345
ss355	2	160819140	160827229
ss364	2	162519455	162509023
ss371	2	25121963	25113787
ss402	2	55533823	55525105
ss405	2	56105254	56094221
ss409	2	56348379	56344786
ss421	2	59569805	59577767
ss439	2	64577637	64581927
ss445	2	67535263	67543823
ss447	2	67614706	67607703
ss453	2	70841946	70846796
ss458	2	77160340	77165312
ss462	2	78913951	78921891
ss465	2	7963579	7974411
ss470	2	83687821	83677707
ss503	3	144684495	144691687
ss505	3	17430081	17432763
ss507	3	18235502	18243716
ss524	3	41953502	41943232
ss525	3	42963477	42958107
ss535	3	53458737	53450492
ss536	3	53616564	53624811
ss541	3	60991937	60984090
ss549	3	70115016	70113404
ss573	4	106488428	106496708
ss581	4	115499635	115507107
ss588	4	127163430	127155271
ss622	4	49165147	49156899
ss623	4	49304506	49312754
ss629	4	51088943	51080961
ss646	4	56259332	56251507
ss647	4	56729720	56737668
ss650	4	59166069	59162169
ss668	4	90421068	90415869
ss669	4	90430977	90424598
ss1181	9	68754775	68764628
ss1182	9	68722493	68714335
ss1186	9	69559106	69564122
ss1204	9	87162707	87156264
ss1224	10	29932464	29924888
ss1244	10	47615354	47622178
ss1248	10	53501690	53493880
ss1254	10	64273141	64278554
ss1259	10	77898120	77906564
ss1260	10	77834589	77840991
ss1263	10	78043151	78051831
ss1264	10	78142110	78133745
ss1284	11	36570871	36577489
ss1288	11	38320818	38329713
ss1291	11	38748083	38755594
ss1293	11	39953037	39947472
ss1297	11	41401109	41409472
ss1301	11	44768980	44774636
ss1329	12	28496925	28505502
ss1334	12	29505296	29500090
ss1335	12	29756546	29761755
ss1339	12	30748621	30743403
ss1401	13	145146779	145151141
ss1402	13	145866907	145872978
ss1406	13	151275361	151284026
ss1410	13	156323812	156315257
ss1411	13	156329581	156315692
ss1416	13	159412739	159402343
ss1426	13	171099331	171091904
ss1472	13	28694473	28685073
ss1491	13	4858668	4865903
ss1509	13	69113303	69106091
ss1548	14	115845419	115838456
ss1553	14	117043403	117034669
ss1567	14	143930291	143925838
ss1578	14	2028560	2034388
ss1586	14	37787786	37792361
ss1602	14	57117365	57113226
ss1611	14	65712859	65699570
ss1615	14	67448499	67455730
ss1623	14	73604698	73611342
ss1649	15	100237597	100246859
ss1662	15	117350577	117343147
ss1667	15	125980416	125988906
ss1673	15	13204349	13202144
ss1688	15	155410633	155405481
ss1707	15	28122024	28128730
ss1734	15	65236271	65227636
ss1746	15	74091967	74084215
ss1752	15	81775510	81778721
ss1760	16	10273998	10278506
ss1761	16	13073345	13082264
ss1778	16	43330995	43339267
ss1784	16	56672786	56661879
ss1791	16	64431989	64437878
ss1806	16	86745527	86750613

ss688	5	108599815	108604258
ss710	5	32349603	32359774
ss725	5	47946400	47938538
ss761	6	111785	114810
ss774	6	112108184	112099226
ss775	6	112106830	112100254
ss788	6	129165720	129171414
ss792	6	133361955	133356323
ss795	6	13752977	13761383
ss797	6	138280270	138289931
ss802	6	146999143	147008780
ss811	6	18873475	18865468
ss832	6	43633817	43631108
ss833	6	43819431	43815063
ss834	6	44095078	44089891
ss842	6	53567761	53559027
ss845	6	53890547	53884860
ss855	6	69772943	69764854
ss874	7	102777448	102769517
ss881	7	112189219	112196743
ss908	7	22829820	22824152
ss940	7	59233125	59225069
ss941	7	60576290	60567608
ss947	7	65641859	65650590
ss981	7	97748570	97757812
ss998	8	123412820	123418627
ss999	8	124048525	124036097
ss1006	8	15578154	15586704
ss1014	8	44237951	44246696
ss1015	8	44716965	44724224
ss1024	8	47014470	47023112
ss1033	8	52125902	52111972
ss1036	8	54301294	54308509
ss1040	8	54563191	54572075
ss1042	8	54872426	54863605
ss1043	8	54821318	54830139
ss1068	8	78532403	78537102
ss1069	8	79184159	79178873
ss1084	9	101122868	101132095
ss1088	9	105677013	105668140
ss1105	9	127452759	127448288
ss1117	9	145809246	145818488
ss1120	9	153033381	153041625
ss1126	9	1954519	1962325
ss1127	9	2006285	1998503
ss1128	9	20377405	20381317
ss1129	9	2083318	2075517
ss1145	9	40592105	40601045
ss1833	17	4053572	4045013
ss1841	17	46915193	46906656
ss1855	18	11855455	11863610
ss1856	18	12085166	12076003
ss1863	18	31426476	31436483
ss1871	18	45559310	45562946
ss1876	18	460853	454479
ss1883	18	533680	529214
ss1929	X	23805143	23798503
ss1959	X	49604464	49601262
ss1961	X	49758214	49765669
ss1968	X	52465132	52473370
ss1969	X	52766971	52758733
ss1975	X	53417953	53409328
ss1976	X	53571662	53562877
ss1985	X	55539535	55547516
ss1994	X	56831905	56841187
ss1998	X	57878868	57873979
ss2002	X	59071736	59062770
ss2009	X	62195630	62203373
ss2028	X	70220328	70229386
ss2030	X	70343608	70334550
ss2033	X	71089360	71097541
ss2038	X	71948191	71937727
ss2051	X	75595630	75600034
ss2069	X	80848767	80840289
ss2075	X	82421018	82409892
ss2078	X	83276403	83285133
ss2079	X	83202921	83200120
ss2085	X	84276659	84268172
ss2092	X	88752006	88758970
ss2099	X	91874109	91867048
ss2104	X	93879217	93887890
ss2106	X	94036848	94028141
ss2112	X	95520352	95528757
ss2113	X	95695773	95688389
ss2117	X	96316717	96324590
ss2118	X	96697842	96705715
ss2124	Y	6721	16672
ss2127	Y	1433768	1427874

musMus_8 _ERV_id	Chr	Start	End
mm631	1	83868369	83860146
mm826	10	25585205	25593428
mm1646	12	20064592	20072882
mm2664	14	40907571	40899181
mm5882	4	51892727	51900951

Supplementary Table 10. Porcine PERVs

	5' end					3' end				
	5' flank Chr	Flank start	Flank end	Flank length	Match starts from	3' flank Chr	Flank start	Flank end	Flank length	Match starts from
PERV A										
3P	Ch1	147,287,040	147,288,299	1,260	1	Ch1	147,296,123	147,294,864	1,260	1
4P	Ch1	294,480,983	294,481,763	1,260	480 - 1260	Ch1	294,489,260	294,488,001	1,260	1
5P*	Ch2	77,159,489	77,160,404	1,260	345 -1260	Ch2	77,167,564	77,166,672	1,260	368 to 1260
15P*	Ch7	112,188,413	112,189,672	1,260	1	Ch7	112,197,277	112,196,018	1,260	1
17P	Ch8	54,510,554	54,511,779	1,260	35 -1260	Ch8	54,570,434	54,569,175	1,260	1
19P	Ch10	77,897,330	77,898,406	1,260	184 - 1260	Ch10	77,906,922	77,905,821	1,260	159 to 1260
22P	Ch12	28,496,119	28,497,378	1,260	1	Ch12	28,506,036	28,504,777	1,260	1
23P	Ch13	151,275,029	151,276,058	1,260	231 - 1260	Ch13	151,284,265	151,283,457	1,260	452 to 1260
33P	ChX	95,519,766	95,520,766	1,260	260 - 1260	ChX	95,529,403	95,528,166	1,260	23 to 1260
PERV B										
1P*	Ch1	42,555,773	42,556,776	1,260	257 - 1260	Ch1	42,562,511	42,561,252	1,260	1 to
8P	Ch3	53,459,559	53,458,300	1,260	1	Ch3	53,449,802	53,451,061	1,260	1 to 1260
10N	Ch4	49,165,755	49,164,759	1,260	215 - 1211	Ch4	49,156,431	49,157,468	1,260	223 to 1260
11N	Ch4	49,165,755	49,164,759	1,260	215 - 1211	Ch4	49,156,431	49,157,468	1,260	223 to 1260
16P	Ch8	15,577,654	15,578,864	1,260	1 - 1211	Ch8	15,587,413	15,586,154	1,260	1 to 1260
18P	Ch9	153,032,559	153,033,818	1,260	1	Ch9	153,042,315	153,041,056	1,260	1 to 1260
24N	Ch13	156,329,810	156,328,810	1,260	260 - 1260	Ch13	156,314,813	156,315,805	1,260	268 to 1260
28N	Ch17	4,054,072	4,053,212	1,260	1 - 861	Ch17	4,052,963	4,053,572	1,260	501 to 1110
30N	Ch17	4,054,072	4,052,881	1,260	1 - 1192	Ch17	4,044,548	4,045,563	1,260	245 to 1260
32N	ChX	83,276,118	83,277,162	1,260	216 - 1260	ChX	83,281,684	83,280,796	1,260	372 to 1260
34N	ChX	128,783,440	128,782,230	1,260	1 - 1211	ChX	128,777,797	128,778,801	1,260	256 to 1260
BETA VIRUS										
B2-1	Ch2	152,131,031	152,130,295	997	261 - 997	Ch2	152,122,064	152,123,060	997	1 to 997
X-1	ChX	53,571,843	53,571,225	1,743	474 - 1092	ChX	53,562,656	53,563,180	1,012	488 to 1012
X-2	ChX	80,849,287	80,848,761	984	1 - 527	ChX	80,840,198	80,840,660	984	432 to 894
X-3	ChX	84,277,179	84,276,650	1,006	1 - 530	ChX	84,267,686	84,268,565	1,006	37 to 916

6. Conserved Synteny and Evolutionary Breakpoints

Identifications of homologous synteny blocks (HSBs)

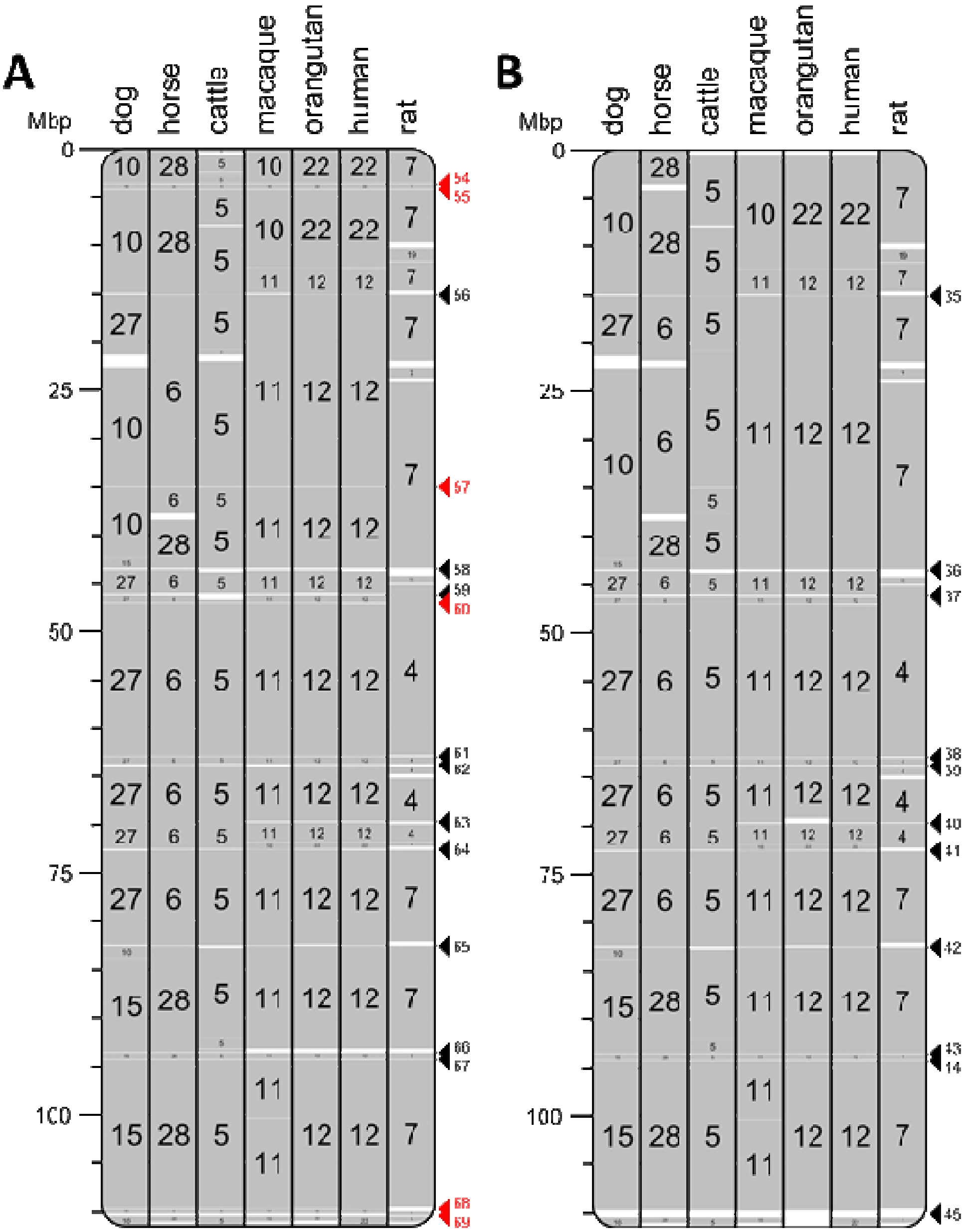
We aligned seven sequenced and assembled mammalian genomes: cattle (UMD 3.0), dog (Cfam2.0), horse (equcab1.0), macaque (mmu2.0), rat (rn4.0), human (hsg37), orang-utan (ponAbe 2.0) with the pig genome (build10.2) using the Satsuma Synteny program (Grabherr *et al*, 2010). The alignments were then cleaned from the overlapping and non-syntenic matches and the homologous synteny blocks (HSBs) were defined using SyntenyTracker (Donthu *et al*, 2009). HSBs were identified using three sets of parameters that allowed the detection of rearrangements that are ≥500kb, ≥300kb, and ≥100kb in the pig genome. In parallel, the same analysis was performed but using the human genome (hsg37) as a reference. Visualization of homologous synteny blocks was performed using the Evolution Highway Comparative Chromosome Browser (http://evolutionhighway.ncsa.uiuc.edu) (Supplementary Fig. 16).

Identification of evolutionary breakpoint regions (EBRs)

Evolutionary breakpoint regions (EBRs) were identified as intervals demarked by two adjacent HSB boundaries on the same chromosome. EBRs were assigned to phylogenetic lineages using the following species relationships: ((pig, cattle), (dog, horse)), (rat, ((human, orang-utan), macaque)). For a reliable classification of EBRs two scores were calculated for each EBR – a *phylogenetic score* and a *gap score*.

The *phylogenetic score* shows if an EBR is present in all species from the expected clade. For example, if an EBR is "pig-specific" and the pig genome was used as a reference for the chromosome comparison, then the highest quality EBR is expected to be present in all species at the same position (phylogenetic score equals 1). If the EBR is not detected in one of the species, then the score will be ~0.86 (1-1/7) given that seven species were aligned with the pig genome. The *gap score* is affected by the number of species in which the EBR is present and whether the EBR detected in one of the genomes overlaps with more than one non-overlapping EBRs in other genomes. For example, if the phylogenetic score equals one and the gap score is less than seven this implies that the EBR present in one of the genomes overlaps with more than one non-overlapping EBRs in other genomes.

If species–specific EBRs are identified using an outgroup genome as a reference, e.g. pig-specific EBRs are detected in the human genome, then a phylogenetic score of one would imply that the EBR is present in only one species. The score would be decreased if the EBR was present in another genome as well, e.g., an overlapping EBR in the pig and mouse genomes has the phylogenetic score of 0.5 implying that it is present in two lineages. The gap score in this case will increase as the number of genomes sharing an EBR increases or if an EBR overlaps with non-overlapping EBRs in other genomes. The algorithm for the EBR classification was implemented as a custom *Perl* script.



Supplementary Fig. 16 Example showing porcine lineage-specific evolutionary breakpoint regions (EBRs) on porcine chromosome 5 at a 300 kb (A) and 500 kb (B) resolution of homologous synteny block (HSB) detection. Red arrows indicate positions of the EBRs detected at the 300 kb resolution but missed at the 500 kb due to a lower resolution of this analysis. A complete set of HSBs defined in our analysis is available from the Evolution Highway comparative chromosome browser (http://evolutionhighway.ncsa.uiuc.edu).

Detection of consensus pig EBRs

These EBR were consistently present in all three HSB datasets or only in the 300 and 100 kb sets (missed in the 500 kb set because of a lower resolution; Supplementary Table 11).

Calculating densities of repetitive elements in EBRs and other parts of the pig genome

A t-test with unequal variances was used to identify repeat families that were unequally distributed in EBRs when compared to the rest of the pig genome (Supplementary Table 12). Local false discovery rate (FDR) critical values (Efron *et al*, 2001) were calculated to control for false positive discovery rate using fdrtool (Strimmer, 2008).

Supplementary Table 11. Pig and primate EBRs in 500kb, 300kb and 100kb resolution HSB sets

Resolution (kb)	Pig EBRs*	Primate EBRs
500	146 (100%)	107 (100%)
300	193 (132%)	127 (119%)
Consensus**	192	NA

*Indicate the number of EBRs present in the porcine and primate lineages and passing stringent thresholds (gap score >2, phylogenetic score >0.86) when defined in the pig and human genomes as a reference, respectively. Percentages indicate a fraction of EBRs identified in the 300 kb resolution set compared to 500 kb resolution (100%). There is an increase in numbers due to a higher resolution of the 300 kb set.

**Consensus EbRs were defined in the pig lineage as those that are consistently present in the sets of 500kb, 300kb and 100kb, or missed only in the 500 kb set because of a lower resolution of this analysis. The consensus EBR set was used for the gene and transposable element enrichment analyses.

Supplementary Table 12. Densities of repetitive element families found to differ significantly in pig- or artiodactyl-specific EBRs.*FDR < 0.05. Repetitive element content is expressed as bp/10kb.

Repeats	Pig EBRs	Other Intervals	Artiodactyl EBRs	Other Intervals
Number of 10kb intervals	2156	257329	210	259275
LINE-L1	1429.41	1332.13	1813.33*	1332.55
LINE-L2	131.89*	256.55	145.20*	255.60
SINE-tRNA-Glu	944.02*	1050.95	1239.61*	1049.91
LTR-ERV1	210.36*	145.15	270.67*	145.59
LTR-ERVL-MaLR	105.41*	160.13	122.13*	159.70
SINE-MIR	116.55*	227.50	102.81*	226.68
DNA-hAT-Charlie	65.35*	111.92	70.33*	111.57
Satellite	300.96*	229.09	368.50	229.57

Selection and fluorescent in situ hybridisation (FISH) mapping of porcine BAC clones

In order to visualise large-scale intra-chromosomal rearrangements between pig and human, whole-chromosome sequences of all chromosomes and their known orthologues (from www.chromhome.org) were aligned using the program GenAlyzer (Choudhuri *et al*, 2004) with default settings. Homologous regions and inversions between chromosome sequences were determined by visual inspection of the GenAlyzer outputs and in order to verify the match information outputs from GenAlyzer, a selection of 71 flanking bacterial artificial chromosomes (BACs) of these EBRs were selected. These BACs were chosen using the porcine BAC end sequence (BES) clone search tool on the Sanger Institute website, (http://www.sanger.ac.uk/resources/downloads/othervertebrates/pig.html) and ordered from the PigE BAC library from ARK-Genomics (http://www.ark-genomics.org/) in order to confirm the *in silico* analysis. Of these 71 BACs, 100% successfully hybridised to the expected porcine chromosomes and chromosomal regions with chromosomal locations being verified by FLPter measurements (fractional length from the p terminus).

Orthologous gene set

We extracted a total of 12,660 pig genes which were annotated by ENSEMBL (build 67), mapped to a chromosome position in the pig genome and which had a single known ortholog in human chromosomes. Because we only considered human-pig 1:1 orthologs, the number of genes used in this analysis is larger than the ortholog set for the 6 mammals that we used for the dN/dS analysis (Section 8). We further filtered this set by excluding those genes located in the non-orthologous positions of the pig and human chromosomes as identified from the whole-genome pig-human SatsumaSynteny (Grabherr et al, 2010) alignment dataset. This dataset had previously been used to build pairwise HSBs between the human and pig genomes. The orthologous positions were identified either by a direct overlap with the pig-to-human sequences alignments, or predicted if a gene was located in between two homologous positions within an HSB as defined by the sequence alignment. We kept those genes that had a single ortholog in the human and pig genomes but were located in a known EBR in the pig genome. As the result of this filtering step 613 genes were removed. To produce a comprehensive set of genes with well-defined orthologous relations between the pig and human genomes we added 116 genes to our dataset that were found in the independent pig genome annotation from NCBI. These additional genes had no physical or name overlap with the annotated pig ENSEMBL gene set and human orthologs located in the homologous positions in the pig and human chromosomes as defined by the sequence-based alignments (see above). We also added 205 genes which had assigned gene names by the NCBI only, were found in homologous positions in human chromosomes confirmed by the whole-genome sequence alignment and had >30% overlap with unnamed pig genes in the ENSEMBL gene set.

The resulting set of 12,368 orthologs between the pig and human genomes was used to build human-pig HSBs with the SyntenyTracker program (Donthu et al., 2009) based only on the gene coordinate information. This led to the detection of 94 genes that were located in unexpected positions within HSBs ("out-of-place") or represented a single gene HSB ("singleton"). We excluded these genes because they are likely to be located in misassembled pig genome intervals and could affect our gene network analysis. At the end, we had the set of 12,274 genes that were used for the gene network analysis.

Detection of porcine bitter taste receptor genes

A total of 105 sequences from known TAS2R genes from cattle, dog, chimp, mouse, human, and pig genomes were collected. A tBLASTn comparison of the genes was performed against the pig chromosomes and unassigned contigs using E-value of e^{-10} as the threshold. All non-overlapping pig sequences that had matches >100 aa within known TAS2R genes were extracted. We added 1,000 bp to the 5' and 3' ends of the extracted sequences. Then we translated all six frames from all the DNA sequences into protein sequences and performed a BLASTp analysis against the NCBI nr database to identify positions of putative TAS2R genes. After detection of the matches we searched the pig sequence for the closest start and stop codons near the longest match from a known TAS2R gene. We considered an identified pig TAS2R gene 'intact' if it encodes for >290 aa, has no frame-shift mutations or premature stop codons. We named new genes found after the family members from other species that had the most significant match in the BLASTp analysis (Supplementary Table 12). In a case when several different pig genes had the most significant match to the same member of TAS2R gene family, we added extensions "A, B, C" at the end of porcine gene names to distinguish between the porcine gene family members.

Supplementary Table 13. Identified intact porcine bitter taste receptor genes.

Gene name	Pig chromosome coordinates	In/near EBR	Annotated by ENSEMBL
TAS2R42	5:63,867,091-63,868,041	YES	NO
TAS2R20	5:63,904,140-63,905,054	YES	NO
TAS2R7A	5:63,940,163-63,941,095	YES	NO
TAS2R7B	5:63,950,624-63,951,541	YES	NO
TAS2R10	5:63,965,446-63,966,375	YES	NO
TAS2R7C	5:63,985,142-63,986,080	YES	NO
TAS2R9	5:63,976,739-63,977,674	YES	YES
TAS2R134	18:5,876,579-5,877,487	NO	NO
TAS2R41	18:7,018,806-7,019,729	YES	YES
TAS2R60	18:7,045,247-7,046,597	YES	YES
TAS2R40	18:7,266,600-7,267,764	YES	YES
TAS2R39	18:7,358,848-7,359,855	NO	YES
TAS2R38	18:8,357,518-8,358,525	NO	NO
TAS2R16	18:25,883,452-25,884,354	NO	NO
TAS2R1	GL893464.1:28,052-29,033	NA	YES
TAS2R3	GL892960.2:34,965-35,915	NA	YES
TAS2R4	GL892960.2:41,686-42,576	NA	YES

Gene network analysis within pig EBRs

The 12,274 pig genes with defined orthologs in the human genome were checked for an overlap with 192 pig-specific EBRs found in pig chromosomes. In total, we detect 1,329 genes that are located within the EBRs or in ±500kb intervals adjacent to the EBR boundaries. To find gene ontology (GO) categories overrepresented in the genes present in pig EBRs, we used human ENSEMBL gene ID.

We used MetaCore GeneGo v.6.9 build 30881 online database to identify GO categories overrepresented in the genes located in/near pig EBRs. We used the complete set of 12,274 orthologous genes as a background for this analysis out of which 12,249 ENSEMBL gene IDs were recognized by MetaCore. Out of 1,329 genes in/near the EBR regions 1,320 were recognized. We later used the Kyoto Encyclopaedia of Genes and Genomes (KEGG) to look into the pathways that were found significantly enriched in the pig EBRs gene set.

Results

The *GO cellular processes* analysis of a stringent set of porcine one-to-one orthologs in the human genome using MetaCore database shows that porcine EBRs and adjacent intervals are enriched for the genes involved in *sensory perception of taste (P<8.9e^{-6};FDR<0.05,* Supplementary Table 14) suggesting that taste phenotypes may be affected by the events associated with genomic rearrangements. A total of ten taste-perception-related genes are present in/near the porcine EBRs (Supplementary Table 15).

Supplementary Table 14. *GO cellular processes* enrichment in pig EBRs.

Processes	P-values	Ratio
Sensory perception of taste	8.9e-6*	11/23
Glutathione metabolic process	8.0e-4	9/25
Sensory perception of bitter taste	1.3e-3	5/9
Midbrain-hindbrain boundary development	1.3e-3	5/9
Regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process	1.3e-3	5/9

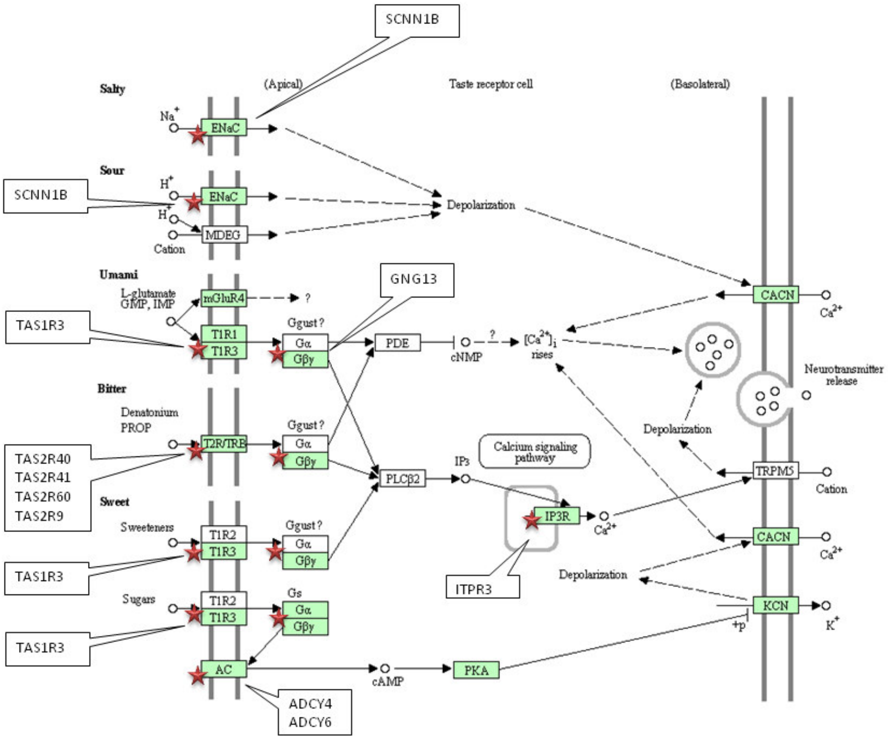
**Sensory perception of taste* was found significant at FDR < 0.05.

SupplementaryTable 15. Genes from taste transduction pathway (KEGG) and taste transduction process (MetaCore) found in/near pig EBRs.

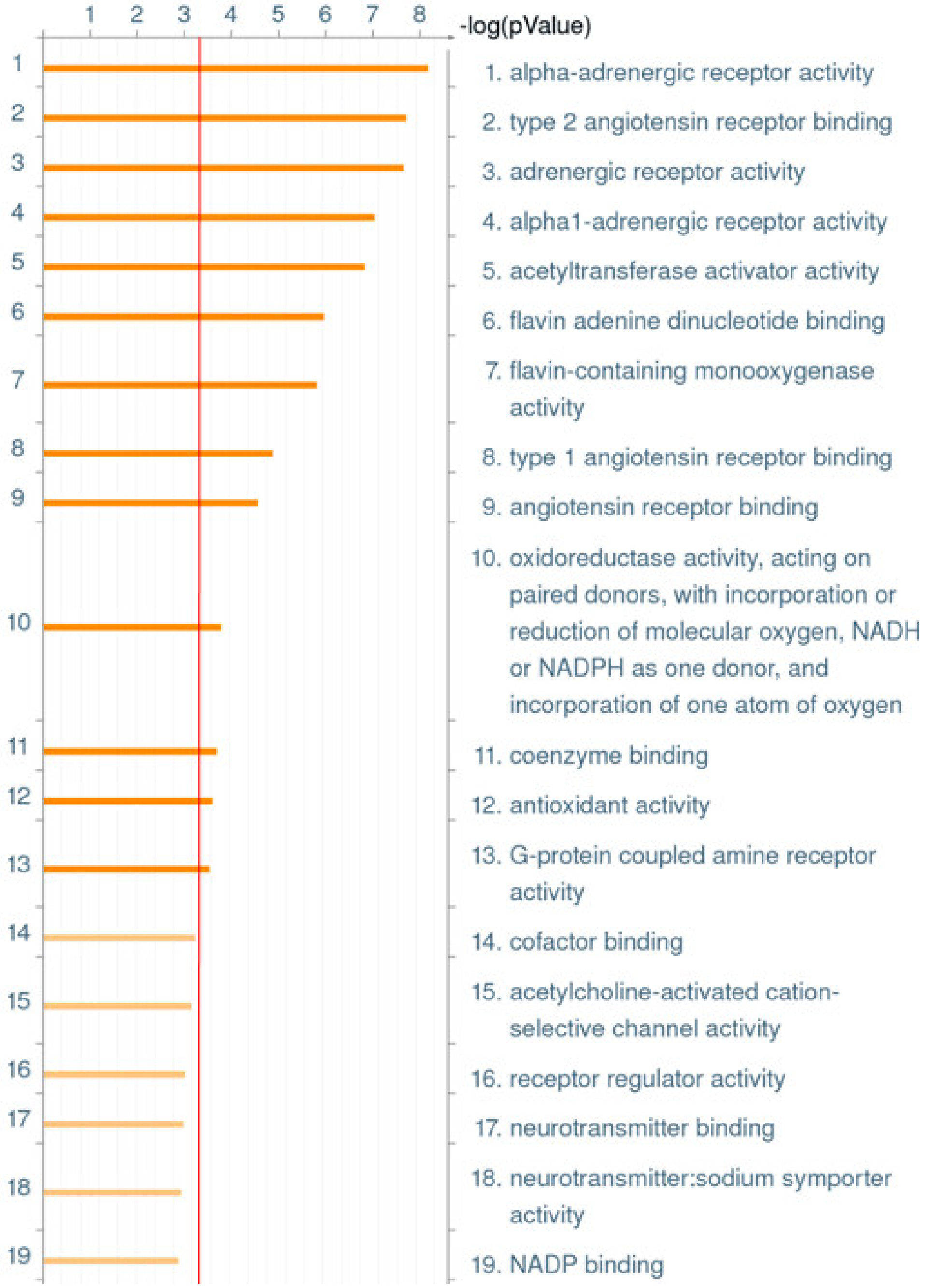
Gene name	Pig chromosome coordinates	Pig EBR coordinates	Database
DBH	1:307,178,223-307,198,524	1:306,934,651-306,985,541	MetaCore
GNG13	3:41,476,392-41,47,6740	3:41,571,689-41,622,736	MetaCore, KEGG
ADCY6	5:15,129,052-15,145,294	5:15,059,839-15,062,939	KEGG
WNT10B	5:15,408,276-15,412,522	5:15,059,839-15,062,939	MetaCore,
TAS2R9	5:63,976,739-63,977,674	5:63,741,431-63,794,981	KEGG
TAS1R3	6:58,111,612-58,115,907	6:57,756,164-57,809,595	MataCore, KEGG
ITPR3	7:34,443,056-34,510,838	7:34,125,342-34,126,061	MataCore, KEGG
ADCY4	7:80,227,590-80,243,075	7:79,938,055-79,942,518	KEGG
SCNN1B	10:309,239-337,906	10:340,718-392,716	MetaCore, KEGG
TAS2R41	18:7,018,806-7,019,729	18:6,766,018-6,823,666	MetaCore, KEGG
TAS2R60	18:7,045,247-7,046,597	18:6,766,018-6,823,666	MetaCore, KEGG
TAS2R40	18:7,266,600-7,267,764	18:6,766,018-6,823,666	MetaCore, KEGG
NPY	18:52,929,674-52,937,567	18:53,339,574-53,398,769	MetaCore

The ten taste-perception-related genes that are present in/near porcine EBRs are highlighted in Supplementary Fig. 17 showing a schematic drawing (KEGG) of the pig taste transduction pathway. In addition to eight *bitter* taste receptors annotated in the pig genome by ENSEMBL we identified 9 additional intact TAS2R genes, bringing the total number of potentially functional TAS2R genes in pigs to 17 (SupplementaryTable 13). Fourteen of the 17 porcine TAS2R genes were assigned to chromosome positions and ten of these were found in clusters near two EBRs on SSC18 (*TAS2R40, TAS2R41, TAS2R60*) and on SSC5 (*TAS2R7A, TAS2R7B, TAS2R7C, TAS2R9, TAS2R10, TAS2R20, TAS2R42*). Some of the porcine taste receptor genes (*TAS1R2, TAS2R1, TAS2R40, TAS2R39*) show an increased foreground to background dN/dS ratio (1.5-1.9) suggesting that these 4 genes have been under relaxed selection. The human genome contains 25 intact bitter taste receptors that originated from primate-specific duplication events (Fisher et al., 2005). The previous studies indicate that pigs normally are not as sensitive to bitter tastes and respond to higher concentrations of bitter compounds than humans (Hellecant and Danilova, 1999; Neslon and Sanregret, 1997), suggesting that pigs are able to use additional food sources that humans cannot. This makes it temping to hypothesize that this feature coupled with a fast growth rates made pigs an attractive species for domestication some 8000 YA. The review of the taste transduction network from the KEGG (Supplementary Fig. 17) shows additional genes affected by the rearrangements and demonstrates that pig genome rearrangements affect *apical* and *taste receptor cell* processes in this network. Together with the results of *GO molecular functions* enrichment analyses (Supplementary Fig. 18) that indicate an overrepresentation of genes related to the *receptor activity* and *binding* categories in the pig

EBRs, our data suggest that chromosomal rearrangements in the Suis lineage have significantly contributed to speciation and adaptation.



Supplementary Fig. 17. Pig taste transduction pathway from the KEGG with red stars indicating the nodes affected by pig genome rearrangements. The genes from the 1:1 ortholog gene set found near/in the EBRs are shown in boxes.



Supplementary Fig 18. Results of *GO molecular function* enrichment analysis in pig EBRs. Categories (1-13) crossing the vertical red threshold line are significant at FDR<0.05.

7 Segmental Duplications

Detection of Segmental Duplication

Segmental duplications (SD; >1kbp and >90% identity) in the pig genome were detected by using whole genome assembly comparison (WGAC) and verified by the whole genome short gun sequencing data (WSSD) using the read depth approach (Sudmant et al. 2010; Alkan et al. 2009). Briefly, for the WGAC, chromosomes were compared pairwise with Blast, using a repeat-masked reference genome, and then duplicated fragments were joined together by re-inserting the previously removed repetitive elements (Bailey et al. 2001). To avoid inclusion of potential false-positive duplications due to the quality of the working draft assembly, SDs with sequence identity > 94% and size >10 kbp were excluded. After the verification by the WSSD method, a total of 1,880 intrachromosomal and 88 interchromosomal high confidence putative SDs were identified as high confident SDs (Supplementary Table 16).

Supplementary Table 16 see excel file "SupTable16.xls"

WSSD to verify SD detected from WGAC method

A read depth method (Sudmant et al. 2010; Alkan et al. 2009) was applied to calculate the copy number of the predicted SD regions. mrsFAST ("Micro-read (substitutions only) fast alignment and search tool" (Hach et al. 2010)) was used to align the whole genome shotgun Illumina reads obtained from Duroc 2-14 (the pig, from which the reference genome sequence is derived) to the repeat masked reference genome (repeat mask information was downloaded from NCBI). We calculated the average read depth for all the predicted SD regions. Next generation sequencing methods result in a bias in the read depth, which is caused by the dissimilar G/C content of different segments of the DNA. We used a similar method to that described in Sudmant et al. (2010) to correct for this GC bias in the sequence data. We analysed the genome in bins of 1 kb and calculated the average read depth across diploid regions for bins of different GC content. We determined the average read depth for each of the different GC classes and used this to obtain a correction factor that was used subsequently to correct the depth of each SD region. After correcting for GC bias, the copy number of all SD regions was calculated based on the depth across the diploid region.

Functional enrichment of genes affected by SDs

Genes that overlap with SDs were identified using the gene annotation from the pig genome (section 2). A total of 263 genes were found to be overlapping with SDs in the pig genome and used as input into DAVID for a gene enrichment and ontology analysis (Huang et al. 2009). Similar to human, mouse, cattle and other mammals, significant gene clusters related to olfaction (41) and immunity (7) were found to be enriched, respectively. Furthermore, 65 genes involved in metal ion transportation were found to be significantly overrepresented in pig SDs.

8 dNdS Analysis

We downloaded the protein and reference mRNA sequences of human, mouse, dog, horse, cow and pig from the ENSEMBL (Hubbard et al. 2002). The 1:1 orthologs of all 6 species were identified by Mestortho (Kim et al. 2008). A total of 9,000 1:1 orthologs for the 6 species was collected. Because we used 1:1 orthologs for all 6 species, the number of genes used in this analysis is smaller than the ortholog data set used for the EBR analysis (Section 6).

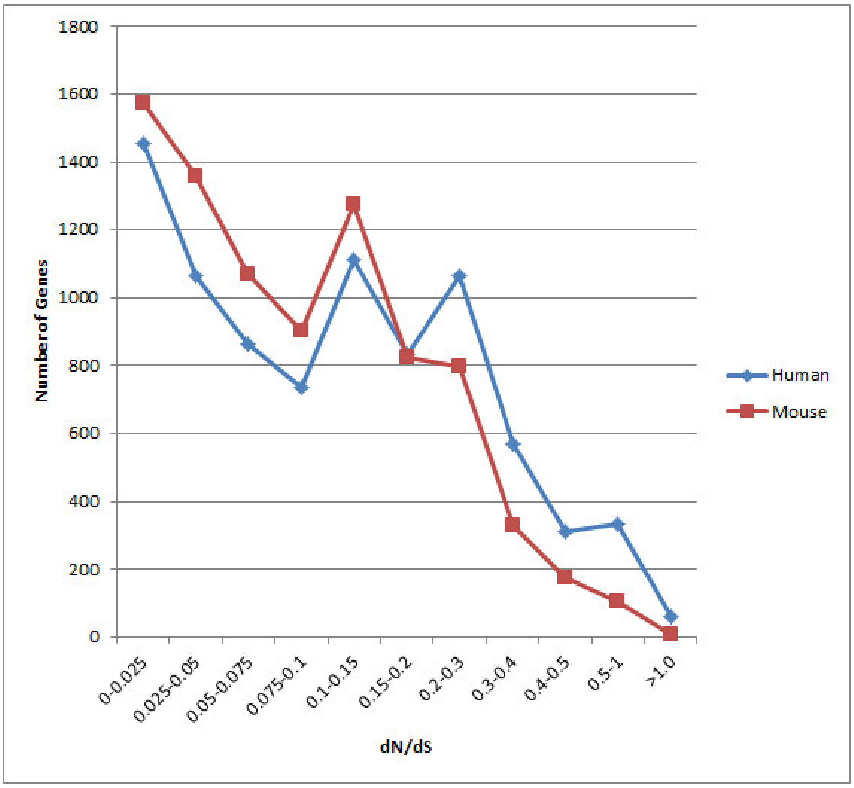
The orthologous gene sets were aligned by prank with its default settings (Löytynoja and Goldman 2005). Poor alignment sites were eliminated using Gblock (Castresana 2005). A maximum likelihood method (codeml of PAML 4 (Yang 2007) was used to estimate to estimate the dN (the rate of non-synonymous substitution), dS (the rate of synonymous substitution) and ω (the ratio of non-synonymous substitutions to the rate of synonymous substitutions) with F3X4 codon frequencies under a branch model (model=2, NSsites=0) and a basic model (model=0, NSsites=0). Orthologs with dS > 3 or ω > 5 were filterd (Castillo-Davis et al 2004) resulting on average in 8,417 orthologs.

The orthologous sets of the 6 mammals were used to identify proteins in each lineage that show accelerated evolution using a branch model implemented in the PAML program. Each lineage was specified in turn as the foreground lineage. Of the orthologous sets, the number of proteins that significantly show accelerated dN/dS ratios in each lineage vary from 84 in the mouse to 311 in the horse lineage. In the pig lineage, 331 proteins showed significant accelerated dN/dS ratios (Supplementary Table 17). The mouse lineage showed the largest average dS value, 0.458, while the other mammals have a relatively similar average dS value ranging from 0.138 (horse) to 0.201 (dog), with an average dS value in the pig lineage of 0.160. When we compare dS values for the different lineages normalised per million years (dS/MY), the same trend is observed, with mouse showing the highest evolutionary rate (0.005038), while ds/MY in the other mammals is considerably lower ranging from 0.001625 to 0.002718 (human 0.001625; horse 0.001662; dog 0.002421; pig 0.002463; cow 0.002718). With regard to the average dN/dS values for each lineage, the human lineage showed the highest value (0.163) whereas the mouse lineage has the lowest (0.116). The average value for dN/dS in the pig lineage is very similar (0.144) to those of the other mammals. The numbers of coding genes per nonoverlapping dN/dS bin in each lineage are shown in Supplementary Fig. 15. Among the six lineages, average dN/dS values differ most between human and mouse. The human lineage has more coding genes with elevated dN/dS ratios in the higher dN/dS bins than the mouse lineage (Supplementary Fig. 19) while the pig lineage has similar values with the others across the different dN/dS bins (Supplementary Fig. 20).

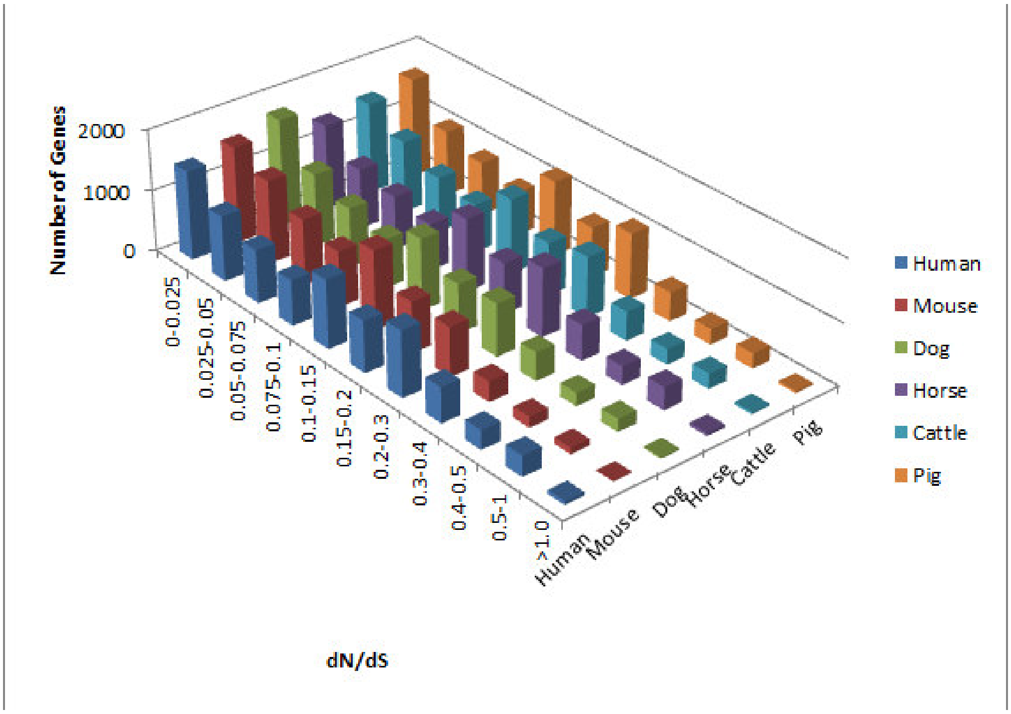
To investigate whether the fast evolving genes in the different lineages were enriched for specific biological processes the coding genes showing increased dN/dS ratios in each lineage were analysed using the DAVID bioinformatics tool (Huang et al., 2009). The enrichment test ($p < 0.1$) for KEGG pathway of the genes, shows that the human and pig lineages show accelerated evolution in ECM-receptor interaction (Supplementary Fig. 21). Considering a branch model for each lineage produces an almost mutually exclusive set of genes between the lineages, suggesting that the adaption of these pathways played an important role during the evolutionary adaptation in these two species.

In order to further evaluate pig as a biomedical model, among the fast evolving genes in the pig lineage, we examined genes known to be associated with human genetic disease (Supplementary Fig. 21). Several disease classes such as psychiatric disorders, cancer, cardiovascular disease, immune diseases and metabolic and neurological disease are highly enriched. Interestingly, recently, a comparative association study between human and pig revealed that many genes related to neurological functions are related to obesity traits in both humans and pigs (Lee et al. 2011).

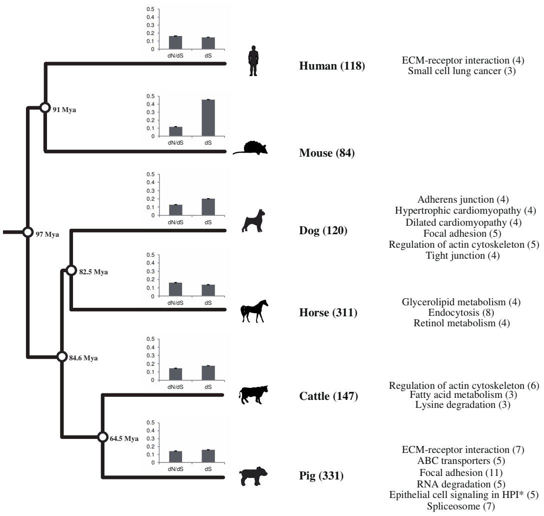
Supplementary Table 17 see excel file “SupTable17.xls”



Supplementary Fig. 19 Distribution of human and mouse genes in the different dN/dS bins



Supplementary Fig. 20 Distribution of genes in the different dN/dS bins for all 6 mammals used in the dN/dS analysis



Disease Class	GENE Symbol
Cardiovascular	ABCA3, ABCC1, JAK2, AP3B1, ROS1, CKM, ITGA1, ITGA2B, ITGA3, NEDD4L, NOS1, NFATC1, ATM, SLC4A5, TGFB2
Psych	CDC42SE2, DDC, DUSP6, GRIN3A, GRIK5, GRN, NOS1, KCNN3, PPP2R2B, RELN, YWHAE, SLC1A3
Developmental	HPS1, JAK2, COL2A1, PAX3, STAT5B, TGFB2
Neurological	ADAM10, ACAD8, BACE2, GRIN3A, GRN, ME2, NOS1, KCNN3, SEMA5A, ATM, SLC11A1, SLC4A3
Cancer	ABCC1, JAK2, RAD50, ALDH1L1, CCND3, FPGS, ITGA3, LTBP1, LHCGR, MAPRE2, NOS1, POLI, ATM, SLC11A1, TGFB2
Reproduction	JAK2, BMPR1B, ITGA2B, LHCGR, SDHA
Chemdependency	RAPGEF3, DDC, NOS1
Renal	NEDD4L, NOS1, XYLT2
Immune	ABCC1, GATA3, COL2A1, EIF2B5, FPGS, IGFBP5, ITGA2B, NOS1, SLC11A1, TGFB2, XYLT2
Metabolic	ABCC1, ABCG8, ACADM, COL2A1, ITGA2B, LRCH1, NOS1, NRF1, PAX8, PFKM, TGFB2

Supplementary Fig. 21 dN/dS and dS values and pathways enriched for genes with elevated dn/dS values in the 6 mammals. The phylogenetic tree of the species is drawn with Timetree (Hedges et al. 2006)

9 Population Genetics (Differentiation and Admixture)

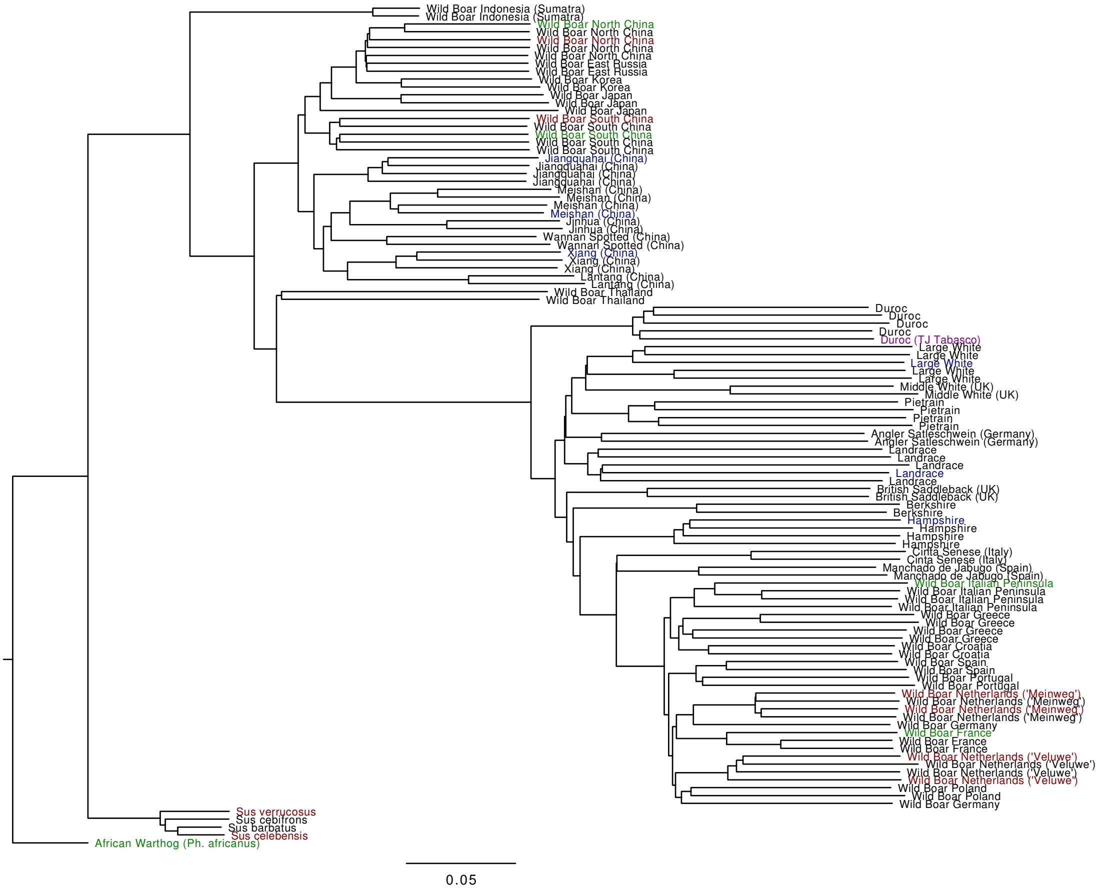
Material and methods:

Selection of individuals

Prior to whole-genome sequencing of the animals used for the present study, 60K single nucleotide polymorphism (SNP) genotype data of candidate animals were compared against a large dataset of 60K SNP genotype data for *Sus scrofa* and related species from all continents with the exception of Antarctica. We have genotyped over 3000 individuals with the Illumina PorcineSNP60 chip (Ramos et al., 2009) and sequenced the D-loop of the mtDNA for all individuals. A representative selection of these individuals is shown in Supplementary Fig. 22. The dendrogram demonstrates that the wild boar sampled for the present study are representative for the geographic extremes of continental Eurasia. In addition, it demonstrates that the domesticated animals used for the population genetic analysis are highly representative for pigs of Europe and China. The analysis was based on 50,492 SNPs from the Illumina PorcineSNP60 chip, that were mapped to the autosomal chromosomes in Sscrofa10.2. The design of the 60K SNP assay was mainly based on SNPs discovered in European pigs, and the very deep genetic divide between European and East-Asian *S. scrofa* that is evident from the current study (see also Megens et al., 2008) is expected to result in a high ascertainment bias. Note that with increasing topological distance to European pigs the branch lengths decrease, which is a clear manifestation of ascertainment bias.

The analysis represents most of the major Eurasian areas as defined by mitochondrial analysis by Larson et al., 2005. For context, *Sus scrofa* from Sumatra was included that represents the 'basal clade' as defined by Larson et al., 2005. Genotyping using the PorcineSNP60 chip on other species in the genus *Sus* and African Warthog would yield relatively high genotyping success (>90%, compared to typically >97% for *Sus scrofa*). Despite assay success, the SNP loci as defined for European pigs would usually not be polymorphic in these species, with only a few percent of SNPs actually found to be polymorphic. Further discussion on the other species in the genus *Sus* are presented elsewhere (Frantz et al. 2012). Note that the two wild boar populations from the Netherlands actually form two distinct populations. Both are more related to the French wild boar than to the Wild boar from the Italian Peninsula. A more detailed analysis of the two Dutch wild boar populations using the PorcineSNP60 chip, is described by Goedbloed et al., 2012.

The European pigs sequenced for this study were all derived from commercial breeds. To demonstrate that these animals are good representatives of the entire breed, animals from at least two distinct populations were selected. Invariably, animals from the same breed cluster together, which shows that despite many generations of selective breeding the breed concept – at least at the genetic level – remains intact as expected (c.f. Megens et al., 2008). Note that the Duroc breed, that also includes TJ Tabasco (Duroc 2-14), tends to cluster basally to all other European or European-derived *Sus scrofa*. This has been observed previously (e.g. Megens et al., 2008). The documented history of the breed is rather sketchy, and despite its clear genetic and historic relationships to European pigs, its precise origin is mostly unknown.



Supplementary Fig. 22. Dendrogram showing the variation that we sampled by re-sequencing, placed in a larger context of animals from the same population. For further context, other animals from different populations were added. The animals used in this study have a blue label if used for the population genetic analysis (Supplementary material section 9), a red label if used for the selective sweep analysis (Supplementary material section 10), or green if used in both. Genotype data from the animal used for the genome assembly (the Duroc sow named 'TJ Tabasco'), was also included with a purple label. A selection of animals from the same population was included, and in addition representatives of other populations of European and East-Asian wild and domestic *Sus scrofa* were included, with black labels. Pigs, wild boar, and outgroups were genotyped using the Illumina PorcineSNP60 chip (Ramos et al., 2009) per the manufacturer's instructions. Pairwise IBS scores were calculated using PLINK v1.07 (Purcell, 2009). Hierarchichal clustering was done using the 'Neighbor' program, which is part of the Phylip phylogenetic analysis package (Felsenstein 2009).

Sequencing, alignment and SNP calling

Genome re-sequencing was targeted at a depth of around 8-10x (for details see Bosse et al., 2012). All sequencing was performed on Illumina HiSeq2000 sequencers. Library construction and re-sequencing of the individual samples was performed with 1-3 µg of genomic DNA according to the Illumina library prepping protocols (Illunima Inc.). The library insert sizes ranged for 300-500 bp and sequencing was performed with the 100 paired-end sequencing kit. Reads were quality trimmed prior to sequence alignment. The trimming strategy involved a 3 bp sliding window, running from 5` to 3`, with sequence data upstream being discarded if the 3-bp window average quality dropped below 13 (i.e. average error probability equal to 0.05). Only sequences 45 bp or more in length were retained. In addition, sequences with mates <45 bp after trimming were also discarded. During trimming, quality scores were recoded to follow the Sanger fastq format to standardize upstream processing.

Sequence alignment was done against the *Sus scrofa* genome, build 10.2, using Mosaik 1.1.0017. We initially used both BWA and Mosaik in our SNP detection pipeline. After evaluation of the false discovery rate in regions of the genome where individuals are homozygous for a single haplotype, it was decided to use Mosaik for our population study. Aligning was done using a hash-size of 15, with maximum of 10 matches retained, and using a 7% maximum mismatch score, for all populations and the outgroup species. Post aligning, alignment files were sorted using the 'Mosaiksort' function, which entails removing ambiguously mapped reads that are either orphaned or fall outside a computed insert-size distribution. Alignment archives were converted to BAM format (Li et al., 2009) using the Mosaiktext function. Manipulations of BAM files, such as merging of alignments archives pertaining the same individual, were done using samtools v. 1.12a (Li et al., 2009).

Variant calling was performed per individual using the 'pileup' function in samtools, and variations were initially filtered to have minimum quality of 50 for indels, and 20 for SNPs. In addition, all variants that had a higher than 3x the average read density estimated from the number of raw sequence reads obtained were also discarded, to remove false positive variant calling originating from off-site mapping as much as possible.

To obtain genotype calls for all the polymorphic sites identified across all individuals, every individual was interrogated for the genotype call for each of the sites found to be polymorphic, including the species-specific differences. Sequence depth, SNP and consensus quality were retrieved for these sites using the samtools pileup function. These *de facto* genotype calls were subsequently filtered based on sequenced depth (minimum sequence depth of 4, and maximum of 2x the average genome-wide depth), where in this case the average sequence depth was established based on the actual sequence depth for each individual separately. Further filtering was done on SNP and consensus quality (in case the individual was homozygous, either SNP or consensus quality > 2, in case the individual was heterozygous, both consensus and SNP quality > 20). All indels were removed. After the filtering, genotype calls were established for a total of 66,668,635 single nucleotide positions in the genome.

For phylogenetic analysis, we identified genomic bins in each sample separately that had an average depth below 2x the genome-wide average depth. We then excluded clusters of 3 SNP in 10 bp and within 3 bp of an indel, in each bin, as these variations are likely to be caused by misalignments (Supplementary Table 18). Finally, we calculated the intersect using BedTools (Quinlan & Hall, 2010), of the genomic bins previously identified for each individual for further analysis (Supplementary Table 19). This resulted in an 11 way alignment with maximum sequence coverage and low false positive variation calling in all our samples.

Supplementary Table 18: Number of filtered SNPs, in autosomal chromosomes and the X chromosome, per individual after filtering for non-uniquely mapping reads.

Sample	Read depth	Fixed SNPs against reference	Heterozygous SNPs
Landrace (LR)	10.4x	2,420,631	2,420,631
Large White (LW)	10.8x	2,616,584	2,254,121
Hampshire (HA)	12.3x	2,875,911	2,004,188
European - NL (WBNL)	11.8x	3,163,655	1,376,164
European – IT (WBIT)	15.1x	3,238,530	1,294,633
Meishan (MS)	9.3x	5,560,909	2,836,716
Xiang (XI)	9.2x	5,481,531	2,696,464
Jiangquhai (JQ)	11.2x	5,124,983	2,750,918
North Chinese (WBNC)	10.7x	4,999,191	3,034,822
South Chinese (WBSC)	10.5x	4,967,382	4,090,363
Phacochoerus Africanus (Pafri)	13.5x	23,000,541	2,159,994

Supplementary Table 19. Summary of fragmented 11-way alignment.

	Total Size	Average Size	Porportion of whole-genome
All	1,232,373,599	2,948	~ 45%
Less than 5kb	626,231,249	1,807	~ 23%
Over 5kb less than 10kb	378,416,929	6,864	~ 14%
Over 10kb	227,725,421	13,864	~ 8%

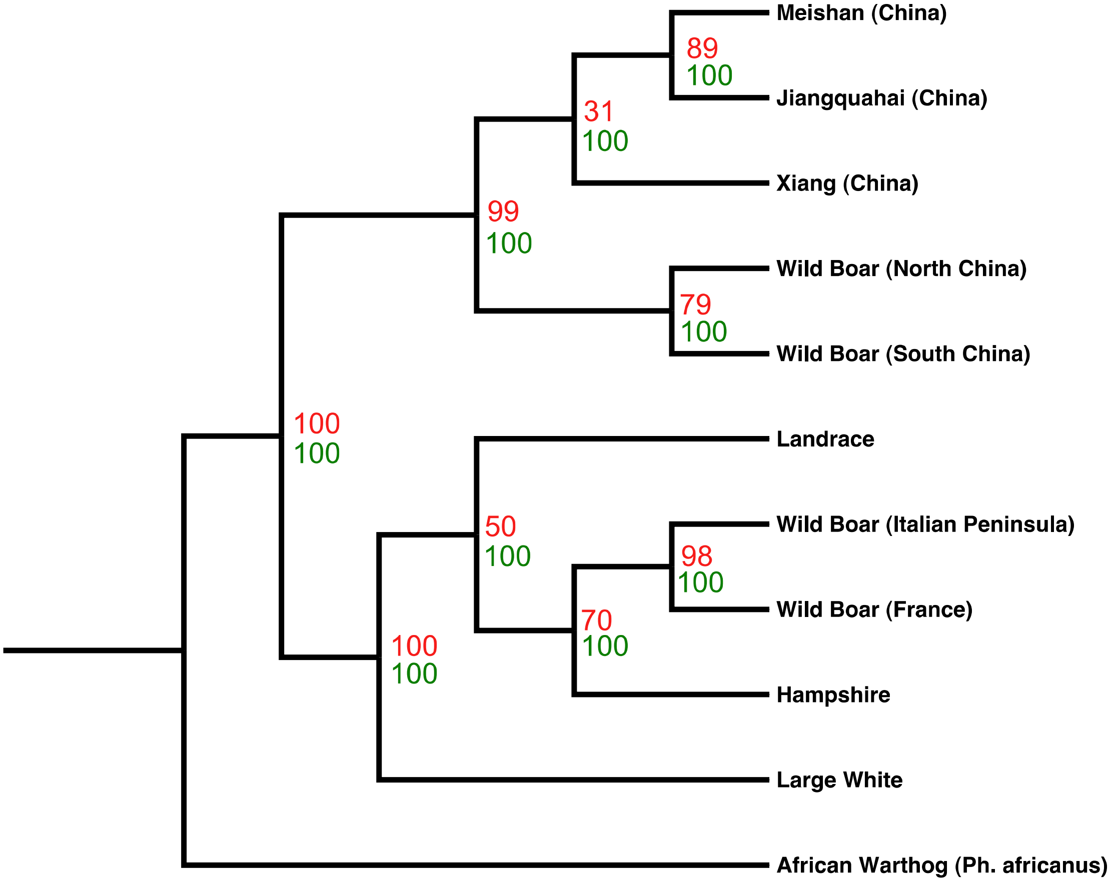
Phylogenomic analysis:

We estimated ML locus trees (bins) using RAxML 7.1.2 (Stamatakis, 2006) with 100 fast bootstrap replicates for each genomic fragment of at least 5 kbp (Supplementary Table 19) to ensure that enough phylogenetic signal was retained in each bin to obtain a reliable tree.

We then built 100 species trees, with one bootstrap replicate from each genomic bin, using STAR (Iu, Ili, & Earl, 2009). Then, we reconstructed a final frequency consensus species tree, from our 100 STAR replicates, using consense from the Phylip package (Felsenstein, 1989).

We computed a concordance factor for each observed clade (Supplementary Table 20). Concordance factor correspond to proportion of each possible clade in the database of bootstrapped single loci trees. Overall the concordance factor supports the main topology (Supplementary Table 20; Supplementary Fig. 23).

Finally, we randomly selected genomic bins (Supplementary Table 19) of minimum 1 kbp to make up 100 non-overlapping alignments of 1Mbp (between 0.99 Mbp and 1.1 Mbp). In each alignment, we fitted a separate GTR+G+I model to each partition (bin) as implemented in RAxML 7.1.2. Thereafter, we ran 100 fast bootstrap replicates for each alignment and a thorough ML search using RAxML 7.1.2. We then constructed 100 frequency consensus trees using one bootstrap replicate from each jackknife replicate and a final frequency consensus tree using all 100 previous consensus using the consense method as implemented in Phylip. This last frequency value was then used as support for species tree (Supplementary Figure 23).



Supplementary Fig 23. Cladogram representing phylogenetic relationship between sequenced pigs. Green values at nodes represent support from STAR analysis, Red values represent support from 100 1 Mbp supermatrices. For abbreviations of the individual pigs, see Supplementary Table 18.

Supplementary Table 20: Concordance factors, that represents the number of time each clade is observed in our database of bootstrapped locus trees. For breed abbreviations, see Supplementary Table 18.

Clade	All	X
WBNL,WBIT	0.2513548567	0.1801490419
HA,LR,LW,WBNL,WBIT	0.2506197065	0.4088360539
JQ,MS	0.2001134871	0.283761533
WBNC,WBSC	0.1531920335	0.081376863
JQ,MS,WBNC,WBSC,XI	0.1426675052	0.1832221434
LR,LW	0.1404972746	0.1234918382
HA,WBNL,WBIT	0.1229877009	0.1228176011
HA,LR,WBNL,WBIT	0.1190078267	0.1264442867
HA,WBIT	0.117712928	0.1180908446
HA,WBNL	0.1176413697	0.1465507452
HA,LR	0.1128497554	0.1160681334
WBNC,XI	0.1124241789	0.1911852378
JQ,WBSC	0.1085276031	0.0786160397
MS,XI	0.108112928	0.1246841732
JQ,XI	0.1080255765	0.1113271824
LR,WBIT	0.1044458421	0.0885805536
HA,LW,WBNL,WBIT	0.1032213836	0.1652164656
LR,WBNL	0.1024497554	0.1042299503
HA,LW	0.1022329839	0.150702626
LR,LW,WBNL,WBIT	0.1019112509	0.0715968772
MS,WBSC	0.0985250874	0.0598722498
WBSC,XI	0.0978923829	0.0651525905
LW,WBIT	0.0941326345	0.0975798439
LW,WBNL	0.0931062194	0.094833215
LR,WBNL,WBIT	0.0905966457	0.0666855926
MS,WBNC	0.0905601677	0.0816252661
JQ,WBNC	0.0901422781	0.1219091554
LW,WBNL,WBIT	0.0898922432	0.0662597587
JQ,MS,XI	0.080093501	0.1511994322
JQ,MS,WBSC	0.0664715584	0.097033357
HA,LR,LW	0.0641870021	0.0823704755
JQ,MS,WBNC,XI	0.0606381551	0.292874379
HA,JQ,LR,LW,MS,WBNL,WBIT,WBSC,XI	0.0585309574	0.0815613911

Heterozygosity analysis

As the average heterozygosity estimates are dependent on the quality and depth of the alignment, SNP calling was executed as previously mentioned with the adjustments that the minimum coverage per position to be included in the analysis was 7x. The total number of heterozygous sites (SNPcount) within an individual that was called in each window of 10 kbp was subsequently corrected by the total number of sites per bin that was sufficiently covered (BINcount), with a minimum coverage of 10%. (Correction factor = 10000/BINcount; Corrected SNPcount = SNPcount*correction factor). Windows that lacked overall coverage of <7X in over 90% of the bin were removed from the analyses due to an increased correction error margin. The average heterozygosity and the proportion of the genome that was included in the heterozygosity analysis are presented in Supplementary Table 21. The distribution of the log2 transformed corrected counts of SNPs per bin are displayed in Fig. 4 of the main text.

Supplementary Table 21: Observed heterozygosity in sequenced individuals. For breed abbreviations, see Supplementary Table 18.

Individual	Average Heterozygosity	Proportion covered
LR	0.00196	0.793
LW	0.00196	0.793
HA	0.00157	0.810
WBIT	0.00116	0.805
WBNL	0.00096	0.809
MS	0.00233	0.807
XI	0.00229	0.804
JQ	0.00228	0.798
WBNC	0.00328	0.800
WBSC	0.00258	0.810

Demographic analysis

We conducted a demographic analysis using a Hidden Markov Model (HMM) approach as implemented in PSMC (Li & Durbin, 2011). PSMC requires diploid consensus sequences. The consensus was generated from the 'pileup' command of SAMtools software package. We applied the same filtering approach as in S1. Then we used the tool 'fq2psmcfa' from the PSMC package to create the input file for the HMM.

We used T_{max}= 20, n = 64 ('4+50*1+4+6'). Plotting the results requires input of generation time and mutation rate. Because there are no convincing data on a different mutation rate in pigs compared to Human we used the default value of $2.5x10^{-8}$ mutation per generation that is the mutation rate in Human. For generation time, we used our best guess and assumed a generation time of 5 years. The results are presented in Fig.2 of the main text.

Admixture analysis – D-statistics:

To detect admixture among our samples we computed D-statistics (Green et al. 2010; Durand et al. 2011). Briefly, with sequence data from one chromosome in 4 different populations P_1, P_2, P_3 and O, where P_1 and P_2 are sister taxa and O is an outgroup, it is possible to infer the state of each allele (derived or ancestral) using the outgroup. Then one can compute the number of derived alleles common between P_1 and P_3 (ABBA count) and between P_2 and P_3 (BABA count). Under the null hypothesis of solely incomplete lineage sorting and no gene flow between P_3 and either P_2 or P_1 we expect a similar count of ABBA and BABA patterns. Under an alternative scenario of gene flow, the count of ABBA must be significantly higher than BABA counts (or vice versa). For a full description of the method please refer to Durand *et al.* (2011). A standard error (SE) of the D-statistics was computed using a Weighted Block Jackknife approach. We divided the genome into N blocks and computed the variance of the statistics over the genome N times leaving each block aside and derived a standard error (SE) using the theory of the Jackknife (For full approach see Green *et al.* Supplementary Online Material 15). We then computed the D-statistics for every possible combination of individuals, using *P. Africanus* as an outgroup. A Bonferroni correction was used to correct for multiple testing by simply multiplying our p-values by the number of D calculations. Because SE may vary greatly depending on block size, we recomputed SE for different block sizes (Supplementary Table 22). Overall these SE estimates were very similar across block sizes. Therefore we used 2 Mb as the block size for further analyses. Finally, we assessed the effect of transition and transversion mutations on D. Overall these resulted in the same outcome (Supplementary Table 23).

Supplementary Table 22: Examples of SE estimation from jackknife analysis using different bin sizes.

P_1, P_2, P_3	2Mbp, D +- SE	5Mbp, D +- SE	10Mbp, D +- SE
WBIT,LW, MS	0.1993+-0.0118	0.1993+-0.0140	0.1993+-0.0167
WBNC MS, LR	0.0676+-0.0074	0.0676+-0.0090	0.0676+-0.0103
WBNC, WBSC, WBIT	-0.0996+-0.0046	-0.0996+-0.0054	-0.0996+- 0.0062
WBNC, WBSC, WBNL	-0.0979+-0.0045	-0.0979+-0.0051	-0.0979+-0.0058

Supplementary Table 23: Examples of the influence of Tv/Ti on D calculation – using 2Mb bins.

P_1, P_2, P_3	Ti, D+-SE	n. ABBA / BABA	Tv, D+-SE	n. ABBA / BABA
WBIT,LW, MS	0.2032+-0.0125	130136 86176	0.1979+-0.0119	349653 234094
WBNC,MS, LR	0.0684+-0.0078	154610 134797	0.0659+-0.0075	416916 365357
WBNC, WBSC, WBIT	-0.1015+-0.0053	119693 146759	-0.0990 +-0.0046	330635 403304
WBNC, WBSC, WBNL	-0.0992+-0.0052	119829 146237	-0.0975+-0.0045	330704 402186

The D statistics are not linearly related to the proportion of admixture (Durand et al. 2011). Computing admixture proportion requires data from a sister taxon to the population that contributed the admixture (an upper bound can be computed with two samples from the same population). In a scenario where we have two sister samples, P1,P2 and P3,P4. If there is an excess of derived lineage from P3 into P2, it is possible to compute the number of common derived alleles between P2 and P4, S(P1,P2,P4). In addition, we can also compute the amount of common derived lineage between P3 and P4, S(P1,P3,P4). The portion of the derived lineage common between P2 and P3 will then behave as if it were a member of P3; hence, S(P1,P2,P4)/S(P1,P3,P4) = f (admixture proportion).

D-statistics interpretation

North Eurasia biogeographic zone.

We found a clear signal for admixture between North Chinese and European populations of wild boars that we interpret as migrations across Eurasia during the later stage of the Pleistocene (Supplementary Table 24). Moreover, this hypothesis is further supported by the high value of concordance factor on the X chromosomes (Supplementary Table 20). The demographic analysis shows that the last glacial maximum (LGM)-induced bottleneck had similar magnitude in Europe and North China (Figure 2, main text). Together, these evidences suggest the existence of another (besides Asian + European) biogeographic zone for pigs, extending across North Eurasia.

Supplementary table 24. Results from D-statistics analysis. First column displays trios involve in D computation. P3 is the population from which we query derived allele into P1 and P2. Second columns represents derived sites considered. The third column displays D value±standard error; and significance level (** p <0.001; * p <0.05; NS non-significant). A positive D value mean admixture in P2, while negative values mean admixture in P1. For breed abbreviations, see Supplementary Table 18.

P1 P2 P3	ABBA BABA	D±SE
WBNC WBSC WBNL	548423 450533	-0.0980±0.0045 **
WBNC WBSC WBIT	550063 450328	-0.0997±0.0047 **
HA WBNL MS	412264 294590	-0.1665±0.0103 **
HA WBIT MS	417559 297334	-0.1682±0.0105 **
LR WBNL MS	470045 313587	-0.1997±0.0102 **
LR WBIT MS	473987 315140	-0.2013±0.0104 **
WBNL LW MS	315903 472494	0.1986±0.0115 **
WBIT LW MS	320270 479789	0.1994±0.0119 **
HA WBNL JQ	407239 302651	-0.1473±0.0116 **
HA WBIT JQ	410441 305295	-0.1469±0.0115 **
LR WBNL JQ	470015 316332	-0.1954±0.0110 **
LR WBIT JQ	473030 318655	-0.1950±0.0111 **
WBNL LW JQ	316752 478129	0.2030±0.0116 **
WBIT LW JQ	322237 484284	0.2009±0.0116 **
HA WBNL XI	411976 275888	-0.1978±0.0108 **
HA WBIT XI	414583 278569	-0.1962±0.0110 **

LR WBNL XI	471840 291850	-0.2357±0.0102 **
LR WBIT XI	473972 294110	-0.2342±0.0100 **
WBNL LW XI	296517 468500	0.2248±0.0098 **
WBIT LW XI	301643 473609	0.2218±0.0103 **
HA WBNL MS	412264 294590	-0.1665±0.0103 **
HA WBIT MS	417559 297334	-0.1682±0.0105 **
LR WBNL MS	470045 313587	-0.1997±0.0102 **
LR WBIT MS	473987 315140	-0.2013±0.0104 **
WBNL LW MS	315903 472494	0.1986±0.0115 **
WBIT LW MS	320270 479789	0.1994±0.0119 **
WBSC XI MS	625181 708687	0.0626±0.0052 **
WBNC XI MS	647153 671457	0.0184±0.0053 NS
WBSC XI JQ	625402 703092	0.0585±0.0054 **
WBNC XI JQ	656865 659913	0.0023±0.0053 NS
MS JQ WBSC	562037 547462	-0.0131±0.0063 NS
MS JQ WBNC	562326 562472	0.0001±0.0069 NS
MS JQ XI	558052 537347	-0.0189±0.0070 NS
JQ XI WBNC	656865 594843	-0.0495±0.0058 **
MS XI WBNC	647153 585888	-0.0497±0.0047 **
JQ XI WBSC	625402 624882	-0.0004±0.0053 NS
MS XI WBSC	625181 610281	-0.0121±0.0046 NS
MS WBSC XI	708755 610282	-0.0747±0.0051 **
WBSC JQ XI	625026 703159	0.0588±0.0059 **
WBNC JQ XI	594947 659974	0.0518±0.0060 **
WBNC MS XI	586000 671512	0.0680±0.0054 **
HA LR WBIT	528296 458122	-0.0711±0.0211 NS
HA LW WBIT	554700 449809	-0.1044±0.0209 **
HA LR WBNL	529654 455864	-0.0749±0.0210 NS
HA LW WBNL	553525 452552	-0.1004±0.0212 **
HA LR JQ	393507 442038	0.0581±0.0136 **
HA LW JQ	401830 458169	0.0655±0.0142 **
HA LR MS	398784 436610	0.0453±0.0132 NS
HA LW MS	409494 448504	0.0455±0.0144 NS

Breed trading.

There was a strong signal for admixture from Asian into European breeds. We found that European domestic breeds such as Landrace and Large White have a significant amount of Asian genetic material (Supplementary Table 24). This admixture is likely to be due to importation of Chinese breeds into Europe (especially UK) at the onset of the 'agricultural' revolution in the late 18th and 19th century. We calculated the admixture fraction of Asian into European breeds as:

$$f_{MS,HA}=\frac{S(WBNL,HA,MS)}{S(WBNL,JQ,MS)}=0.33$$

$$f_{MS,LR}=\frac{S(WBNL,LR,MS)}{S(WBNL,JQ,MS)}=0.38$$

$$f_{MS,LW}=\frac{S(WBNL,LW\ ,MS)}{S(WBNL,JQ,MS)}=0.38$$

Within Asia.

The difficulty of building a phylogenetic tree for the breeds within Europe and China is puzzling. Many aspects such as incomplete lineage sorting (ILS), breed trading, multiple domestication origin, husbandry practices and biogeographic pattern could explain these results. In Asia, the clustering of breeds illustrated in Figure 23, appears to be much more complex. The Meishan and Jiangquhai pigs do not share significantly more derived alleles with Chinese wild boras or Xiang (Supplementary Table 24), which suggests that these two breeds have a common wild ancestor and did not undergo admixture since their separation. This is not surprising as these breeds are from very similar geographic areas. Thus, bootstrap values and concordance factors can be explained solely by ILS. However, this is not the case for the Xiang breed. We found that North Chinese wild boar derived alleles match Meishan or Jiangquhai significantly more often than Xiang (Supplementary Table 24). This is expected as Meishan and Jiangquhai are from Northern regions of China. Surprisingly, Xiang do not share significantly more derived allele with Southern Chinese wild boar than with Northern wild boar, MS and JQ, which is in agreement with the South Chinese origin of Xiang. In addition, the Xiang's derived alleles are found significantly more often in Jiangquhai and Meishan than in both Northern and Southern Chinese wild boar (Supplementary Table 24). Lastly, we found Jiangquhai derived alleles 6% more often in Southern Chinese wild boar than Xiang. This pattern highlights the composite origin of the Xiang breed. Such a finding can be explained by complex breed trading and admixture with local wild boar populations within China and / or multiple origins of domesticated pigs in China. Thus, our analyses do not allow us to distinguish between these hypotheses. Further studies, that capture more genetic diversity within Asia may be able to provide an answer to this question.

Within Europe

In Europe, the clustering of breeds and wild boar seems even more complex. Breeds do not form a monophyletic group as one would expect if they shared a common wild ancestor. For example, the derived lineages from Dutch wild boar match the Hampshire lineage 10% more often than the Large White lineage (Supplementary Table 24), thus, supporting our phylogeny (Supplementary Fig. 23). As we showed above, these breeds have different degrees of Asian genetic material, imported during the agricultural revolution. Thus, this may solely explain the paraphyly of European breeds. Under this null hypothesis we expect that alleles coming from Asian admixture will influence the topology and D value. Let us suppose that the Large White has more Asian derived alleles than Hampshire, when querying the alleles of the Dutch wild boar in these breeds as D(HA,LW,WBNL), we expect that the excess of Asian alleles in Large White compared to Hampshire influences our calculation, thus making Hampshire closer to the Dutch wild boar. To test this hypothesis we re-computed the Dstat pulling out every derived allele common between (LW, MS+JQ) and (HA, MS+JQ), thus, minimizing the Asian influence in our European calculations. We found that Asian alleles had a very minor influence on our calculation. For D(HA, LW, WBNL) we considered 1,006,195 derived sites. When removing sites where either Large White or Hampshire matched Meishan and Jiangquhai rather than the Dutch wild boar, this number fell to 1,006,172. Moreover, removing those sites did not influence our estimated of D (D(HA, LW, WBNL) = 0.1003). In addition, we found that both D(LW,HA,JQ) and D(LW,HA,MS) show an excess of match between Large White and Jiangquhai or Meishan, however this was not significant using Meishan (Supplementary Table 24). Thus we hypothesise that Asian admixture is not solely responsible for the paraphyly of European breeds. Other factors such as husbandry practices (see main text) and / or multiple domestication origin in Europe probably played an important role.

10 Selective Sweep Analysis European – Asian Wild Boar.

Sequencing, alignment and SNP calling

We sequenced the genomes of 10 individual wild boar (Supplementary Table 25) at a targeted depth of around 8-10x (for details see Bosse et al., 2012). These included the 4 wild boar used in the population genetic analysis described under section 9 and in addition: 1 wild boar from South China, 1 from North China, 1 from France and 3 from the Netherlands. Illumina (v. 1.3-1.7) formatted fastq files, with sequences between 60 (Illumina GA2, part of the data for *Sus verrucosus* and *Sus celebensis*) and 100 bp (Illumina GA2 and HiSeq2000). Sequence alignment and SNP calling is described under section 9. These 10 wild boar are representative of the wild boar populations in Asia and Europe (see section 9, Supplementary Fig. 22).

Supplementary Table 25: Number of filtered SNPs, in autosomal chromosomes and the X chromosome, per individual after filtering for non-uniquely mapping reads.

Sample	Read depth	Fixed SNPs against reference	Heterozygous SNPs
European - NL (WBNL1)	11.8x	3,163,655	1,376,164
European - NL (WBNL2)	9.9x	2,951,757	1,367321
European - NL (WBNL3)	5.7x	2,487,211	635,079
European - NL (WBNL4)	8.4x	2,952,957	1,034,760
European - FR (WBFR1)	9.6x	3,015,677	1,378,353
European – IT (WBIT1)	15.1x	3,238,530	1,294,633
North Chinese (WBNC1)	4.9x	3,441,649	1,347,806
North Chinese (WBNC2)	10.7x	4,999,191	3,034,822
South Chinese (WBSC1)	4.3x	3,828,708	1,946,662
South Chinese (WBSC2)	10.5x	4,967,382	4,090,363
Sus celebensis	30.0x	14,757,346	3,388,064
Sus verrucosus	13.4x	15,386,605	490,861
Phacochoerus Africanus	13.5x	23,000,541	2,159,994

Derived allele frequencies at variable site in European wild boar

In total we have identified 17,210,760 variable sites within the 10 sequenced wild boars, 2,212,288 of which still show variation in both the European and Asian wild boar populations. The majority of these shared SNPs represent, polymorphisms that were already present in the ancestral population before the European - Asian wild boar split, some 1 million years ago. The remaining 14,998,472 variable sites represent a mixture of sites that are fixed in one or both of the populations. These variable sites therefore represent:

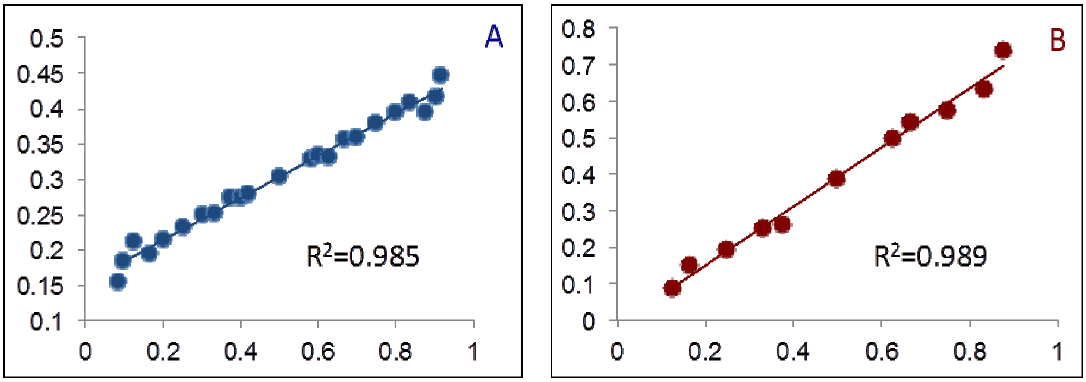
(1) Sites that occur at different frequencies in the European and Asian populations, for which the minor allele was not sampled in the small collection of European or Asian individuals sequenced in our study

(2) Positions that got fixed for one of the alleles in either the European and/or Asian populations after the split of European/Asian split.

(3) New mutations that arose in either the European or Asian population after the European/Asian split.

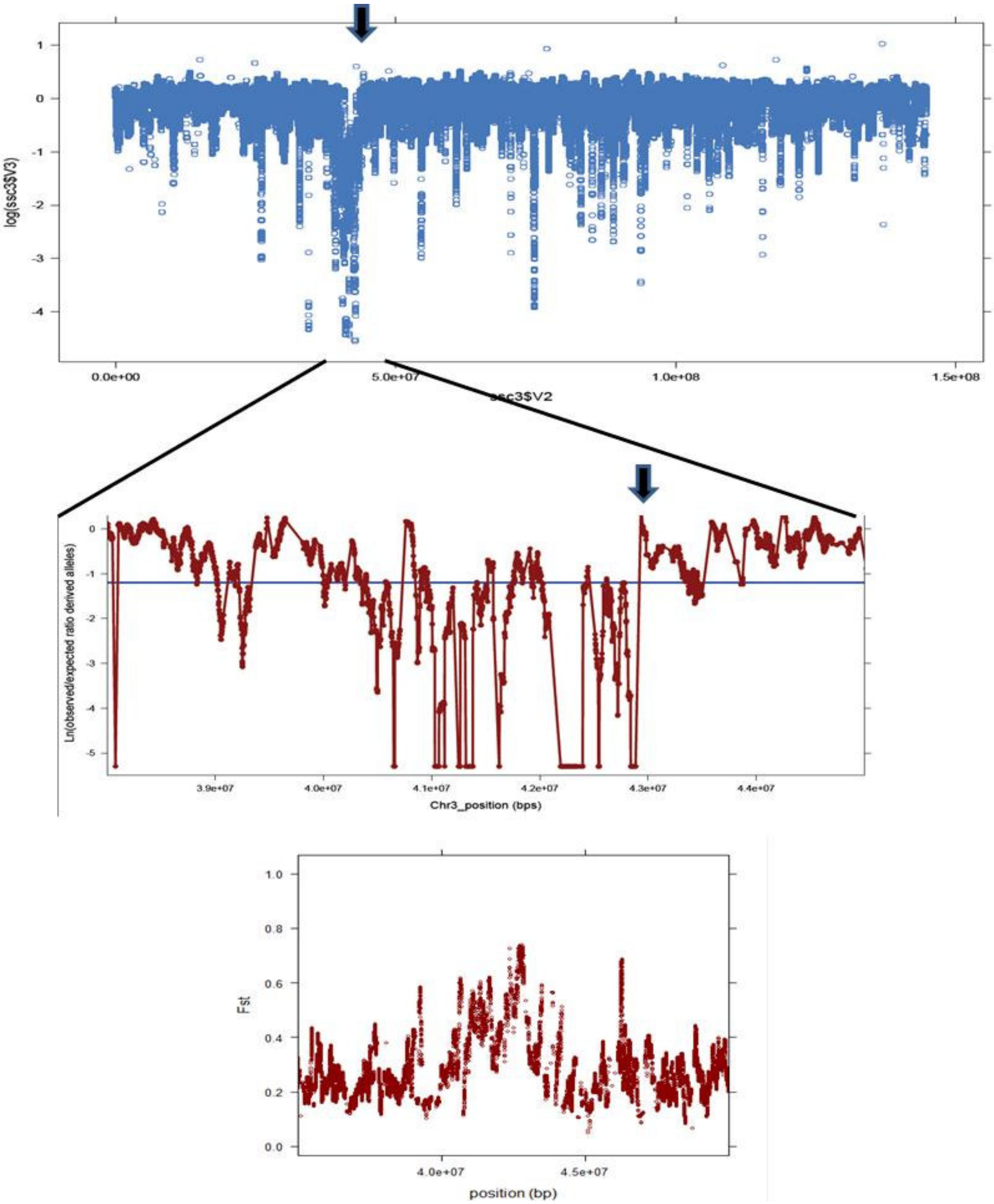
On average, a genomic region not under selection will share a similar number of variable sites and these sites thereby provide a measure for selective sweeps that occurred after the split between the European and Asian wild boar. For regions that have become fixed after this split due to a selective sweep, such shared polymorphisms will have been lost.

We performed a search to identify such regions using an approach similar to that used previously for the analysis of selective sweeps in the human genome after the split between human and Neanderthal some 400,000 years ago (Green et al. 2010). To be able to determine the most likely ancestral state for the observed alleles at the variable sites, we used whole genome sequence data from 3 other Suids: *Phacoechorus africanus*, *Sus verrucosus* and Sus *celebensis* (Frantz et al. 2012). If one of the *Sus scrofa* alleles was observed in the African Warthog (*Phacochoerus africanus*), *Sus verrucosus* or *Sus celebensis*, in that order respectively, that particular allele was assumed to represent the ancestral state. We then calculated the derived allele frequency at all 17,210,760 variable sites and selected only those sites where the derived allele frequency in the European wild boars was greater than 0. We then calculated the derived allele frequencies at these sites in the Asian wild boars. As expected, due to the incomplete lineage sorting of SNPs present before the European Asian split, the derived allele frequencies in both populations are highly correlated (R^2=0.985; Supplementary Fig. 24A). Similar results were obtained if the analysis was done starting from the Asian population for sites where the derived allele frequency in the Asian population is greater than 0 (.R^2=0.989; Supplementary Fig. 24B).



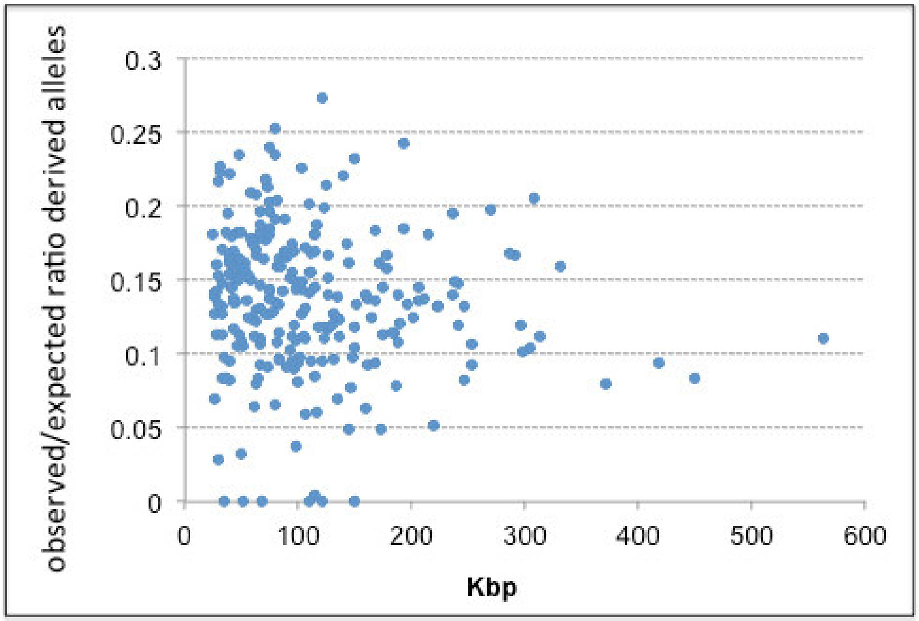
Supplementary Fig. 24. (A) Correlation between observed derived allele frequency between European and Asian populations at sites where the derived allele frequency in the European population is >0. X-axis: derived allele frequency in European. Y-axis: derived allele frequency in Asian. (B) Correlation between observed derived allele frequency between European and Asian populations at sites where the derived allele frequency in the Asian population is >0. X-axis: derived allele frequency in Asian. Y-axis: derived allele frequency in European

For positions in the European population with an observed derived allele frequency > 0, this correlation was used to calculate the expected derived allele frequency at that position in the Asian population and this was compared to the actual observed derived allele frequency at that position. We plotted the log-transformed value of the ratio for the observed/expected derived allele frequency using a sliding window at a bin size of 50,000 bp (Supplementary Fig. 25).



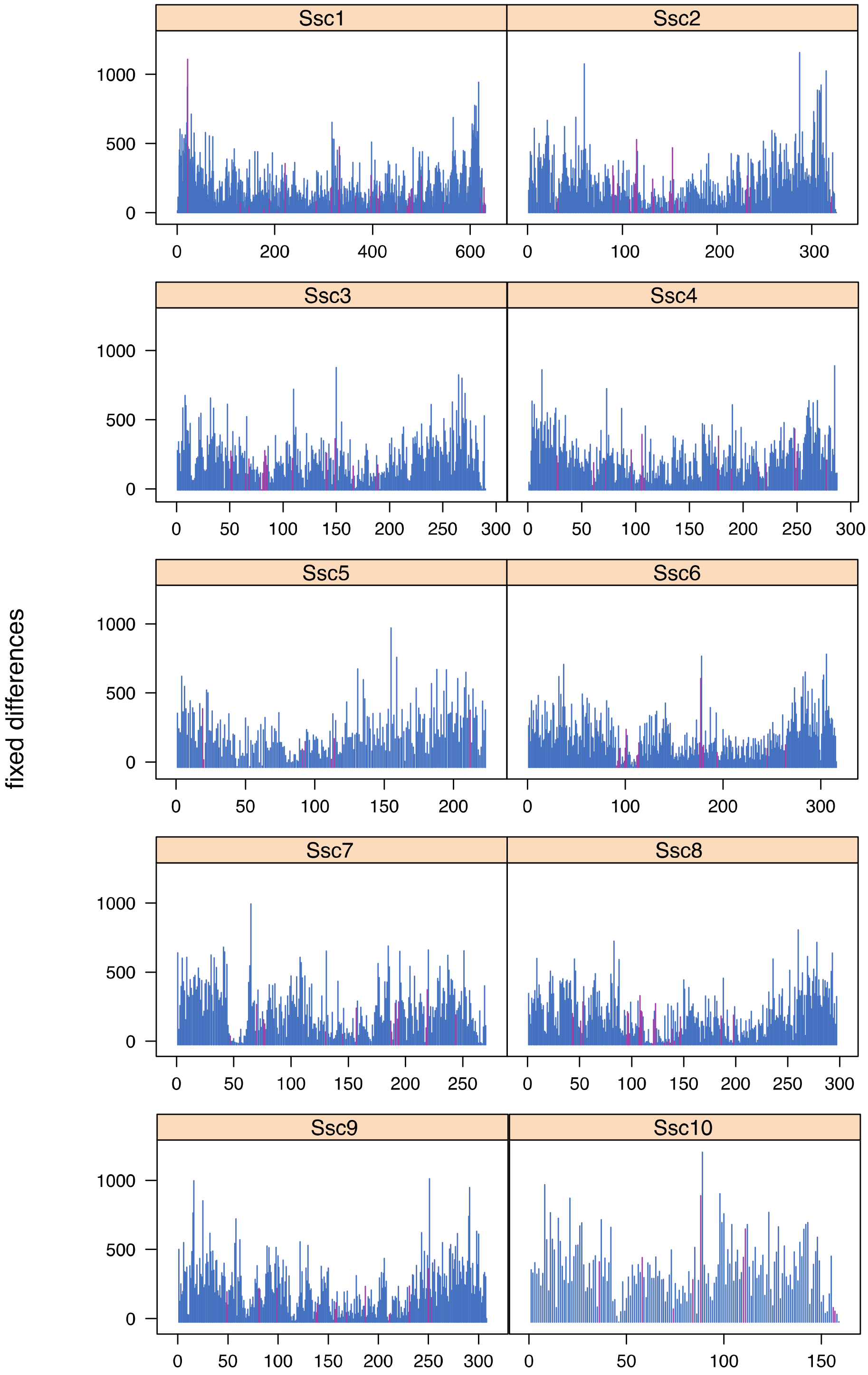
Supplementary Fig. 25. Selective sweep regions on chromosome 3. Plot of the log transformed ratio of the observed/expected derived allele frequency in the Asian wild boars at positions where the derived allele frequency in the European wild boars is >0. The most likely position of the centromere is indicated by an arrow. The zoomed in region shows the Selective sweep regions between positions 39-42 MB on this chromosome. Lowe panel shows the Fst values for the same region.

Finally, we identified regions in the genome where the frequency of derived alleles is significantly lower than the expected genome average using the following thresholds: The size of the region should be larger than 25,000 bp and the ratio for the observed/expected derived allele frequency should be lower than 0.3. This resulted in a total of 251 putative selective sweep regions, with an average size of 111,269 bp and harbouring 365 annotated protein-coding genes (Supplementary Table 26, Supplementary Fig. 26 and 27).

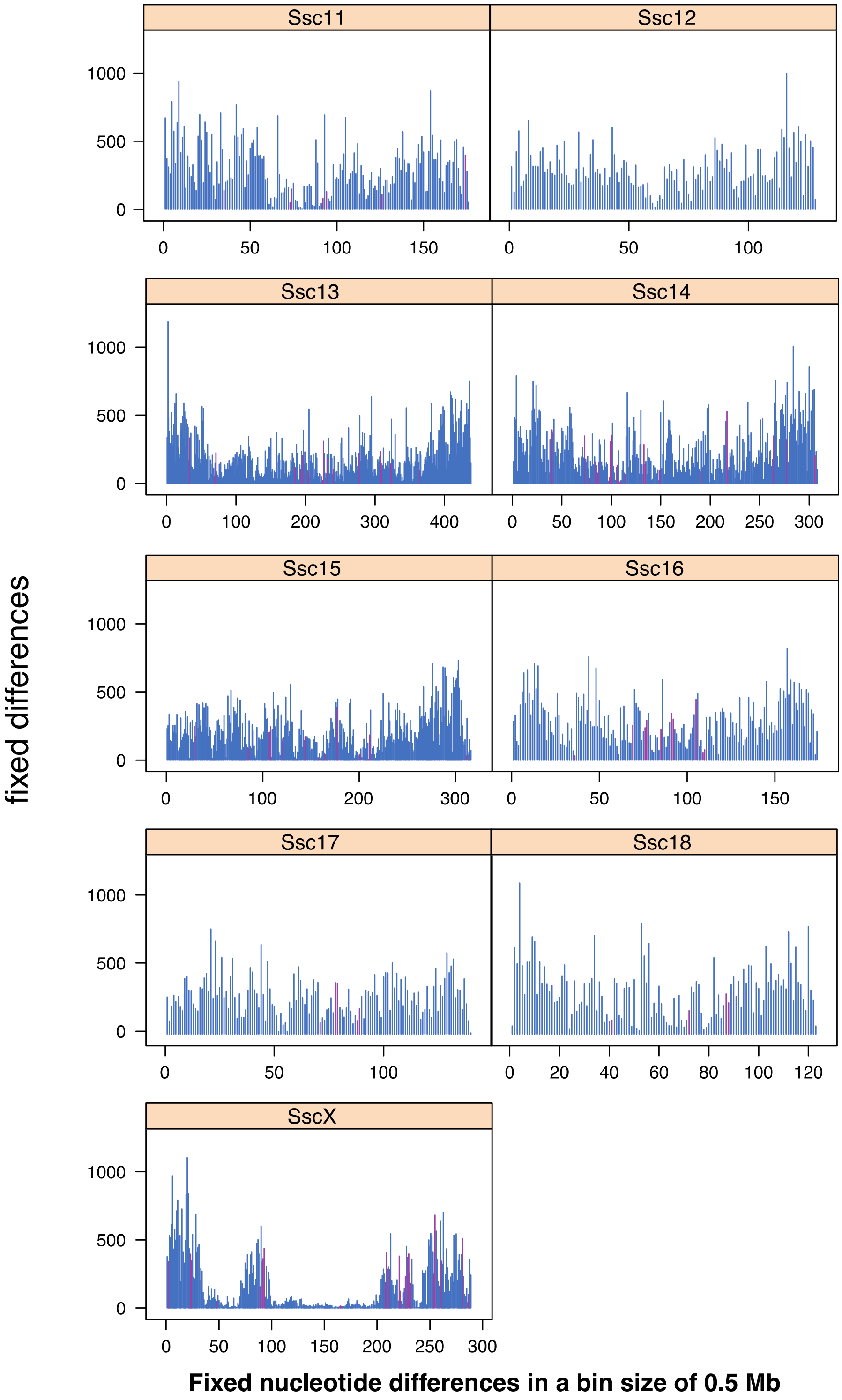


Supplementary Fig. 26 Candidate selective sweep regions. All regions shown are >25,000 bps with an observed/expected derived allele frequency ration of <0.3. x-axis: Size in bp. y-axis: ratio of observed/expected derived allele frequency

Supplementary Table 26 see excel file “SupTable26.xls”.



Fixed nucleotide differences in a bin size of 0.5 Mb



Supplementary Fig. 27. Differentiation between European and Asian wild boar. Fixed nucleotide differences between European and Asian wild boar was counted within bins of 0.5 Mb. The X-axis shows the bin number on each chromosome and the Y-axis shows the number of fixed nucleotide differences that were observed. The 0.5 Mb bins that overlap with the selective sweep regions are shown in purple.

GO enrichment analysis of genes within selective sweep regions

To investigate whether specific classes of genes or pathways, have been under strong selection in the European and/or Asian wild boar, we performed a GO term enrichment analysis on the genes located within the identified selective sweep regions. We decided to exclude the large region on chromosomes 3 between 39-43 Mb from our analysis. This region shows very low levels of recombination and harbours a large number (90) of different genes. Because actual selection within this region might have been on only a small number of the genes located within this region, including all genes from this region in our analysis would result in substantial background noise obscuring potential true signals from the other sweep regions. Because of the relatively low number of GO terms connected to the remaining 275 genes, we decided to use GO terms connected to the orthologous human genes. The GO term enrichment analysis was performed using R package Gostats (Falcon et al. 2007). The conditional algorithm was used for the hypergeometric test. The gene annotation package for the GOstats analysis was built using R package AnnotationDbi (Pages et al. 2008). Mapping of porcine Ensembl gene IDs and other genomic information (e.g. entrezgene) was performed using the R package biomaRt (Durick et al. 2005). The results show a clear enrichment with genes that are involved in RNA splicing and RNA processing (Supplementary Table 27).

The selective sweep regions demonstrated an enrichment of genes that are involved in RNA processing and regulation (*CELF1, DGCR14, RBM5, SCAF1, CELF6, HNRNPA1, HNRNPM, WDR83, RBM39, SF3A1 SYMPK, TXNL4A*). This is surprising, as this is a class of genes involved in highly conserved processes and the genes in general are highly conserved. A recent simulation study by Pavlides et al. (2012) emphasizes that great care has to be taken when interpreting the results of GO enrichment analysis, and that such analysis can easily result in storytelling. While surprising to find overrepresentation of genes involved in RNA processing and splicing in regions of the genome that exhibited strong selection during the split of the European and Asian breeds, additional observations in our study and from the literature support a hypothesis in which changes in alternative splicing might be a way for a species to rapidly adapt to a new environment:

(1) In our dNdS analysis (see supplementary information section 8) we observed accelerated evolution of genes involved in splicing and RNA degradation

(2) Two genes (*SCAF1* and *HNRPA1*) located within a selective sweep region encode for different protein variants fixed in the European and Asian wild boar.

(3) The 12 genes involved in RNA processing and regulation are located in 12 different selective sweep regions and therefore comply with the assumption of independence, underlying the GO enrichment analysis.

(4) Many metazoan splicing factors belong to large gene families and rapid evolution of

specific splicing factors after gene duplication has recently been described in Drosophila (Taliaferro et al., 2011).

(5) Genetic variation in splicing factors within populations has been described. Although generally being identified in relation to specific diseases, this highlights that this type of variation can affect phenotypic variation (Garcia-Blanco et al, 2004).

Alternative splicing is often tissue specific and large differences are observed in the alternative transcripts in different species. It is assumed that alternative splicing is influenced by a large number of factors and that subtle changes in the relative abundance of these factors within different tissues can have profound effects on the use of the weaker splice sites in particular (Ast 2004, Kalsotra and Cooper 2011). This raises the intriguing hypothesis that small changes in the expression of such factors might be a way for rapid evolution within a species.

Supplementary Table 27 Gene enrichment analysis on putative selective sweep regions excluding the 10 regions on chromosome 3 between 39-43 Mb.

GOBPID	Pvalue	Odds Ratio	Nr	Size	Term
GO:0000377	6.1E-04	5.32	7	85	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GO:0000398	6.1E-04	5.32	7	85	nuclear mRNA splicing, via spliceosome
GO:0000375	1.1E-03	4.76	7	94	RNA splicing, via transesterification reactions
GO:0000380	1.2E-03	17.52	3	13	alternative nuclear mRNA splicing, via spliceosome
GO:0008380	1.5E-03	2.95	12	256	RNA splicing
GO:0043517	1.7E-03	58.13	2	4	positive regulation of DNA damage response, signal transduction by p53 class mediator
GO:0046838	1.7E-03	58.13	2	4	phosphorylated carbohydrate dephosphorylation
GO:0046855	1.7E-03	58.13	2	4	inositol phosphate dephosphorylation
GO:0000245	1.8E-03	8.68	4	31	spliceosome assembly
GO:0006397	2.3E-03	2.79	12	270	mRNA processing
GO:0080010	2.3E-03	13.47	3	16	regulation of oxygen and reactive oxygen species metabolic process
GO:0046856	2.8E-03	38.75	2	5	phosphoinositide dephosphorylation
GO:0010812	3.3E-03	11.67	3	18	negative regulation of cell-substrate adhesion
GO:0016071	4.0E-03	2.48	13	327	mRNA metabolic process
GO:0046839	4.2E-03	29.06	2	6	phospholipid dephosphorylation
GO:0006139	5.7E-03	1.48	76	3458	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO:0010310	5.8E-03	23.25	2	7	regulation of hydrogen peroxide metabolic process
GO:0043949	5.8E-03	23.25	2	7	regulation of cAMP-mediated signaling
GO:0006396	6.5E-03	2.08	17	510	RNA processing
GO:0032088	6.8E-03	8.75	3	23	negative regulation of NF-kappaB transcription factor activity
GO:0060391	7.6E-03	19.37	2	8	positive regulation of SMAD protein nuclear translocation

GO:0007162	8.3E-03	5.44	4	47	negative regulation of cell adhesion
GO:0043484	9.6E-03	7.61	3	26	regulation of RNA splicing
GO:0000381	9.7E-03	16.6	2	9	regulation of alternative nuclear mRNA splicing, via spliceosome
GO:0048821	9.7E-03	16.6	2	9	erythrocyte development
GO:0060390	9.7E-03	16.6	2	9	regulation of SMAD protein nuclear translocation

Fixed protein variants within the selective sweep regions

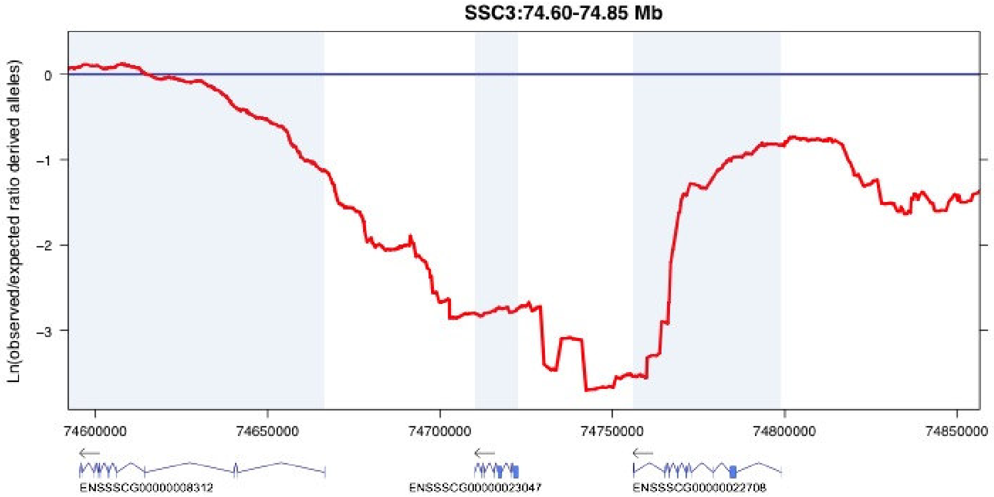
Within the complete genome we have identified 1,706 non-synonymous mutations in 1,191 genes for which different alleles are fixed in European and Asian wild boars, which based on the total number of annotated protein coding genes in the current build (21,578) amounts to 5.5%. A GO enrichment analysis of these 1,191 genes indicates that this group of genes is enriched for genes involved in sensory perception, immunity and host defence (Supplementary Table 28).

Supplementary Table 28 Gene enrichment analysis on the 1,191 genes for which different alleles that code for different protein variants are fixed in the European and Asian wild boars.

GOBPID	Pvalue	Odds Ratio	Nr	Size	Term
GO:0006956	2.49E-05	6.39	10	35	complement activation
GO:0002541	5.45E-05	5.70	10	38	activation of plasma proteins involved in acute inflammatory response
GO:0006957	8.71E-05	11.93	6	14	complement activation, alternative pathway
GO:0006959	1.59E-04	3.78	13	68	humoral immune response
GO:0051604	4.12E-04	2.91	16	104	protein maturation
GO:0016485	4.66E-04	3.00	15	95	protein processing
GO:0006958	4.84E-04	6.19	7	25	complement activation, classical pathway
GO:0002455	5.60E-04	5.10	8	33	humoral immune response mediated by circulating immunoglobulin
GO:0019835	6.09E-04	7.34	6	19	cytolysis
GO:0006954	1.46E-03	1.86	33	320	inflammatory response
GO:0051605	1.49E-03	3.04	12	75	protein maturation by peptide bond cleavage
GO:0007586	1.65E-03	2.84	13	86	digestion
GO:0031648	1.97E-03	10.57	4	10	protein destabilization
GO:0007156	3.10E-03	2.51	14	103	homophilic cell adhesion
GO:0006952	3.21E-03	1.57	50	566	defense response
GO:0090195	3.55E-03	Inf	2	2	chemokine secretion
GO:0090196	3.55E-03	Inf	2	2	regulation of chemokine secretion
GO:0090197	3.55E-03	Inf	2	2	positive regulation of chemokine secretion
GO:0060592	3.68E-03	15.84	3	6	mammary gland formation
GO:0007608	3.70E-03	1.75	32	326	sensory perception of smell
GO:0007600	4.31E-03	1.51	56	659	sensory perception
GO:0071109	6.16E-03	11.88	3	7	superior temporal gyrus development

GO:0007606	6.50E-03	1.66	33	352	sensory perception of chemical stimulus
GO:0002526	8.26E-03	2.39	12	92	acute inflammatory response
GO:0006281	8.84E-03	1.74	26	266	DNA repair
GO:0008360	8.87E-03	3.03	8	50	regulation of cell shape
GO:0001867	9.42E-03	9.50	3	8	complement activation, lectin pathway

As expected, the number of genes fixed for different protein coding variants is significantly higher within the identified selective sweep regions (12.6%) where 46 of the 365 genes code for different protein variants in the European and Asian wild boars (shown in red in Supplementary Table 26). For example in the gene ENSSSCG00000023047 homologous to the *ZNF638* gene (Supplementary Figure 28) a non-synonymous mutation in exon 1 of the gene results in an Ala230Pro substitution. Individual genome sequence data from European and Chinese domestic breeds show the same contrast with alanine found at this position in the European breeds and proline in the Chinese breeds. Similarly as for the ERI2 variants, Chinese haplotypes coding for the proline variant are segregating in Pietrain. We have further analysed the Ala230Pro substitution using PolyPhen2 and this amino acid change is likely to affect the function of the protein. Another striking example is the *ABCA3* gene, which, like, the ZNF638 gene, is located on SSC3 close to the centromere. Whereas the *ZNF638* gene codes for a transcription factor that has recently been shown to be a novel regulator of adipogenesis and early adipocyte differentiation (Meruvu et al, 2011), the *ABCA3* gene is proposed to play a role in lipid organization and lipid transport. Genes involved in fat deposition and metabolism are of great interest in animal breeding and numerous quantitative trait loci (QTL) involved in backfat thickness have been identified based on crosses between European and Asian breeds (see PigQTLdb: http://www.animalgenome.org/cgi-bin/QTLdb/SS/index). QTL for backfat thickness have been mapped to the regions on SSC3 containing the *ZNF638* and *ABCA3* genes, using crosses between Chinese and European pigs (Beeckmann et al, 2003; de Koning et al. 2003).



Supplementary Fig. 28. Putative selective sweep regions at the ZNF638 like genes ENSSSCG00000023047 and ENSSSCG000022708 on SSC3. The gene ENSSSCG00000023047 is homologous to exons 1-10 of the human ZNF638 gene whereas the gene ENSSSCG000022708 is homologous to exons 22-18 of the human ZNF638 gene. The current pig genome assembly probably is missing the sequences for exons 11-21, resulting the gene to be split into two gene models.

11 Pig as a biomedical model

Because the pig is used extensively as a biomedical model, we were interested to address a number of basic questions: (1) Do specific pig proteins exist that carry amino acid substitutions that in human are implicated in specific disease phenotypes? (2) Can we identify pigs that harbour loss of function (LoF) mutations in genes that in human are implicated in specific disease phenotypes? With regard to the first question, we both examined such variants in the genome of Duroc 2-14 (aka TJTabasco), the pig whose genome was used for the reference sequence and in the genome of an additional 48 individuals representing a broad collection of breeds and wild boar from Europe and Asia. The identification of potential disease causing alleles (non-synonymous as well as LoF) that are segregating in the pig population, offer the possibility to develop additional new biomedical models to study specific diseases relevant for humans.

To address question 1, we first examined the porcine proteins in the reference sequence which is from a Duroc pig. Although the Duroc breed was developed in the United States it can be considered to be European in the context of the European-Asian comparisons discussed above. We aligned all pig and human orthologous proteins using blastp in order to identify all amino acids that are different between these species at orthologous positions. This resulted in 1,393,618 amino acid positions where the pig and human proteins differ. We then examined whether any of these amino acid differences resulted in an amino acid substitution that in humans is implicated in the development of a disease. Towards this end, amino acid substitutions that are implicated in the development of human disease were downloaded from the online mendelian inheritance in man (OMIM) database at NCBI and from Ensembl Variation 66 using BioMart. To avoid the inclusion of amino acid substitutions at poorly annotated segments within the current pig gene models, we only included positions where the local sequence identity in a 20 amino acid window around the variant between the human and pig protein was equal or higher than 50%. The resulting 112 substitutions are shown in Supplementary Table 29. All 48 pigs sequenced have the same amino acid at this position as the reference genome.

Potentially more interesting variants are those for which both a normal and a potential disease-causing variant are segregating in pig populations. We applied the filtering settings for SNP calling as described in the "Selective Sweep Analysis European – Asian Wild Boar" section above and also only considered positions where the local sequence identity in a 20 amino acid window around the variant between the human and pig protein was equal or higher than 50%. This resulted in a total of 32,548 non-synonymous mutations and 269 nonsense mutations. Among the non-synonymous mutations, there were 6 that resulted in the same amino acid substitution that has been implicated to play a role in human disease (Supplementary Table 30). Except for the *HBB* and *AGTR2* variants, several homozygous individuals are seen for the mutations in the *MYBPC3*, *DDC*, *CTNS* and *RPS19* genes, suggesting that these variants have a minor effect on the phenotype in pigs. Of the 157 nonsense mutations, 11 are linked to a human disease phenotype (Supplementary table 31).

Supplementary Table 30 [a] Difference fixed between European and Asian pigs

Chr	Pos	alleles	Gene	aa_change	Ref allele count	Alt allele count	OMIM	Hs aa change
2	16456620	T/C	MYBPC3	V59A	67	15	600958	T59A
9	5641101	A/G	HBB	L115P	95	1	141900	L115P
9	150228713	T/C	DDC	K147R [a]	74	26	107930	S147R
12	51600079	A/G	CTNS	I42V	17	87	606272	V42I
17	425162	C/T	RPS19	R62W	62	20	603474	R62W
X	109824242	C/T	AGTR2	R323Q	87	1	300034	R324Q

Supplementary Table 29 see excel file "SupTable29.xls"

Supplementary Table 31 see excel file "SupTable31.xls"

12 References:

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061–1067.

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

Anderson SI, Lopez-Corrales NL, Gorick B, Archibald AL, (2000) A large fragment porcine genomic library resource in a BAC vector. *Mammalian Genome* 11:811-814.

Ast G. (2004) How did alternative splicing evolve? Nature Reviews Genetics 5, 773-782

Bakewell MA, Shi P, Zhang J. (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. PNAS 104: 7489-7494

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 11:1005-17.

Bakewell MA, Shi P, Zhang J. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci U S A. 2007 May 1;104(18):7489-94Barreiro, L.B., Quintana-Murci, L. (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. Nature Reviews Genetics 11:17-30

Beck S (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics 60:1-18.

Beeckmann P, Schroffel Jr. J, Moser G, Bartenschlager H, Reiner G, Geldermann H (2003) Linkage and QTL mapping for Sus scrofa Chromosome 3. J Anim Breed Genet. 120:20-27

Benson G. (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573-580 http://tandem.bu.edu/trf/trf.html

Binns RM, Duncan IA, Powis SJ, Hutchings A, Butcher GW. (1992) Subsets of null and gamma delta T-cell receptor+ T lymphocytes in the blood of young pigs identified by specific monoclonal antibodies. Immunology. 77:219-227

Bosse M, Megens, H-J, Madsen O, Paudel Y, Frantz L et al. (2012). Regions of homozygosity in the porcine genome: Consequence of demography and the recombination landscape. PLoS genetics in press

Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science 324:522-528

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol. 268:78-94

Busset J, Cabau C, Meslin C, Pascal G (2011) PhyleasProg: a user-oriented web server for wide evolutionary analyses. Nucleic Acids Res. 39:W479-85

Castillo-Davis C, Kondrashov F, Hartl D, Kulathinal R (2004) The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome research 14: 802-811

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17: 540-552

Chain PSG, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Woolam A, Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, Detter JC, 2009. Genome project standards in a new era of sequencing. Science 326: 236-237

Choudhuri JV, Schleiermacher C, Kurtz S, Giegerich R. (2004) GenAlyzer: interactive visualization of sequence similarities between entire genomes. Bioinformatics 20:1964-1965

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. (2004) The Ensembl automatic gene annotation system. Genome Res. 14:942-950

Davuluri RV, Grosse I, Zhang MQ (2001) Computational identification of promoters and first exons in the human genome. Nat Genet. 29:412-417

Dawson, H. (2011) Comparative assessment of the pig, mouse, and human genomes; A structural and functional analysis of genes involved in immunity. Pp. 323-342. The Minipig in Biomedical Research. P. A. McAnulty, Editor-in-Chief, A, Dayan, K. H. Hastings, N-C. Ganderup, (eds.) CRC Press (Taylor & Francis Group, LLC). Boca Raton, FL.

de Koning DJ, Pong-Wong R, Varona L, Evans GJ, Giuffra E, Sanchez A, Plastow G, Noguera JL, Andersson L, Haley CS (2003) Full pedigree quantitative trait locus analysis in commercial pigs using variance components. J Anim Sci. 81:2155-2163

Dong D, Jones G, Zhang S. (2009) Dynamic evolution of bitter taste receptor genes in vertebrates. BMC Evolutionary Biology, 9:12

Donthu R, Lewin HA, Larkin DM. (2009) SyntenyTracker: a tool for defining homologous synteny blocks using radiation hybrid maps and whole-genome sequence. BMC Res Notes 2:148.

Down TA, Hubbard TJ (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res 12:458-461 http://www.sanger.ac.uk/resources/software/eponine/

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. Molecular biology and evolution 28:2239-2252

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21:3439-3440

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113

Edgar RC, Myers EW (2005) PILER: identification and classification of genomics repeats. Bioinformatics 21; Suppl. 1: i152-i158

Efron B, Tibshirani R, Storey J, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* **96**, 1151

Ellinghaus D. Kurtz S, Willhoeft, U, (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Genomics 9: 18.

ENA http://www.ebi.ac.uk/ena/

Eguchi-Ogawa T, Toki D, Uenishi H. (2009) Genomic structure of the whole D-J-C clusters and the upstream region coding V segments of the TRB locus in pig. Dev Comp Immunol. 33:1111-9

Eguchi-Ogawa T, Wertz N, Sun XZ, Puimi F, Uenishi H, Wells K, Chardon P, Tobin GJ, Butler JE. (2010) Antibody repertoire development in fetal and neonatal piglets. XI. The relationship of variable heavy chain gene usage and the genomic organization of the variable heavy chain locus. J Immunol. 184(:3734-42.

Eyras E, Caccamo M, Curwen V, Clamp M (2004) ESTGenes: alternative splicing from ESTs in Ensembl. Genome Res. 14:976-987

Fahrenkrug SC, Rohrer GA, Freking BA, Smith TP, Osoegawa K, Shu CL, Catanese JJ, de Jong PJ (2001) A porcine BAC library with tenfold genome coverage: a resource for physical and genetic map integration. Mammalian genome 12: 472-474

Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. Bioinformatics 23:257-258

Fan B, Onteru SK, Du ZQ, Garrick DJ, Stalder KJ, Rothschild MF (2011) Genome-wide association study identifies Loci for body composition and structural soundness traits in pigs. PLoS One. 6:e14726

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, *5*(2), 163-166. doi:10.1111/j.1096-0031.1989.tb00562.x

Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D, (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biology and Evolution 1: 205-220

Fischer A, Gilad Y, Man O, and Pääbo S. *(*2005) Evolution of bitter taste receptors in humans and apes Mol Biol Evol 22: 432-436

Frantz L, Shraiber J, Madsen O, Megens H-J, Semiadi G, Li N, Crooijmans RPMA, Archibald, AL, M., Slatkin M, Schook LB, Larson, G., & Groenen, M. A. M. (2012). Genomic sequencing provides fine scale inference of evolutionary history. *Submitted*

Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 26:407-415.

Garcia-Blanco MA, Baraniak AP and Lasda EL Alternative splicing in disease and therapy. *Nature biotechnology* 22, 535-546 (2004)

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A (2011) Rfam: Wikipedia, clans and the decimal release. *Nucleic Acids Research* 39:D141-145

Goedbloed DJ, Megens HJ, Van Hooft P, Herrero-Medrano JM, Lutz W, Alexandri P, Crooijmans RP, Groenen M, Van Wieren SE, Ydenberg RC, Prins HH. (2012) Genome-wide single nucleotide polymorphism analysis reveals recent genetic introgression from domestic pigs into Northwest European wild boar populations. Mol Ecol. June 26. [Epub ahead of print]

Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R (2010) A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res. 38 Suppl:W695-699.

Grabherr MG, Russell P, Meyer M, Mauceli E, Alföldi J, Di Palma F, Lindblad-Toh K. (2010) Genome-wide synteny through highly sensitive sequence alignment: Satsuma. Bioinformatics. 26:1145-1151

Green R.E, Krause J, Briggs A.W, Maricic T, Stenzel U, Kircher M, Patterson N, et al. (2010) A draft sequence of the Neandertal genome. Science *328*:710-722

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. Nucleic Acids Research 31:p439-441

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. NAR 2006 34(Database Issue):D140-D144

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Meth 7: 576–577

Hedges SB, Dudley J, Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22: 2971-2972

Hellecant G, Danilova V. (1999) Taste in domestic pig, Sus scrofa. J. Anim. Physiol. a. Anim. Nutr. 82: 8–24

Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, Forbes S, Gilbert JG, Halls K, Harrow JL, Hart E, Howe K, Jackson DK, Palmer S, Roberts AN, Sims S, Stewart CA, Traherne JA, Trevanion S, Wilming L, Rogers J, de Jong PJ, Elliott JF, Sawcer S, Todd JA, Trowsdale J, Beck S (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics 60:1-18

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 4:44-57

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, et al. (2002) The Ensembl genome database project. Nucleic acids research 30: 38-41

Humphray SJ, Scott CE, Clark R, Marron B, Bender C, Camm N, Davis J, Jenks A, Noon A, Patel M, Sehra H, Yang F, Rogatcheva MB, Milan D, Chardon P, Rohrer G, Nonneman D, de Jong P, Meyers SN, Archibald A, Beever JE, Schook LB, Rogers J (2007) A high utility integrated map of the pig genome. Genome Biol. 8:R139.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 44:226-232

Iu, L. I. L., Ili, L. Y. U., & Earl, D. E. K. P. (2009). Estimating Species Phylogenies Using Coalescence Times among Sequences, 58:468-477

Juhling F, Morl M, Hartmann RK, Sprinzl M, Stadler PF, Putz J (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Research 37:D159-D162

Jurka J, Kapitonov VV, Pavlicke A, Klonowski P, Kohany O, Walichiewicz J, (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenetics and Genome Research 110: 462-467

Kalsotra A and Cooper TA (2011) Functional consequences of developmentally regulated alternative splicing. Nature Reviews genetics 11:715-729

Kent WJ (2002) BLAT-the BLAST-like alignment tool. Genome Res. 12:656-664

Kim KM, Sung S, Caetano-Anollés G, Han JY, Kim H (2008) An approach of orthology detection from homologous sequences under minimum evolution. Nucleic acids research 36: e110-e110

Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research* 39:D152-D157

Kusza S, Flori L, Gao Y, Teillaud A, Hu R, Lemonnier G, Bosze Z, Bourneuf E, Vincent-Naulleau S, Rogel-Gaillard C (2011) Transcription specificity of the class Ib genes SLA-6, SLA-7 and SLA-8 of the swine major histocompatibility complex and comparison with class Ia genes. Anim Genet. 42:510-20

Kuzio J, Tatusov R, and Lipman DJ (2006) Dust. Unpublished but briefly described in: Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. Journal of Computational Biology 13:1028-1040

Lee KT, Byun MJ, Kang KS, Park EW, Lee SH, et al. (2011) Neuronal Genes for Subcutaneous Fat Thickness in Human and Pig Are Identified by Local Genomic Sequencing and Combined SNP Association Study. Plos One 6

Lestrade L, Weber MJ (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research*, 34:D158-D162

Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME (2002) Apollo: a sequence annotation editor. Genome Biol. 3:RESEARCH0082.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20:265–272

Li H, Durbin, R (2011) Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493-496

Loveland JE, Gilbert JG, Griffiths E, Harrow JL (2012) Community gene annotation in practice. Database (Oxford). 2012:bas009

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:955-64.Pages H, Carlson M, Falcon S and Li N (2008) AnnotationDbi: Annotation Database Interface. R package version 1.4.0

Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America 102:10557-10562

Megens HJ, Crooijmans RP, San Cristobal M, Hui X, Li N, Groenen MA. (2008) Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples:

differences in microsatellite variation between two areas of domestication. Genet Sel Evol. 40:103-28

Mosig AF, Guofeng MF, Stadler BM FAU - Stadler P, Stadler PF (2007) Evolution of the vertebrate Y RNA cluster. Theory Biosci. 126:9-14

Nelson SL, Sanregret JD (1997) Response of pigs to bitter-tasting compounds. Chem Senses. 22:129-32

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: A fast search method for large DNA databases. Genome Res 11:1725-9.

Onteru SK, Fan B, Du ZQ, Garrick DJ, Stalder KJ, Rothschild MF (2012) A whole-genome association study for pig reproductive traits. Anim Genet. 43:18-26

Pages H, Carlson M, Falcon S, Li N (2008) AnnotationDbi: Annotation Database Interface. R package version 1.4.0 2008

Pavlides P, Jensen JD, Stephan W and Stamatakis A (2012) A critical assessment of storytelling: Gen ontology categories and the importance of validating genomic scans. Mole. Biol. Evol. [Epub ahead of print] PubMed PMID: 22617950

Pasman Y, Saini SS, Smith E, Kaushik AK (2010) Organization and genomic complexity of bovine lambda-light chain gene locus. Vet Immunol Immunopathol. 135:306-13

Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M (2004) The Ensembl analysis pipeline. Genome Res. 14:934-941

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. In: 13th International Conference on Intelligent Systems for Molecular Biology, Detroit, Michigan, USA, 25-29 June 2005. Bioinformatics 21: Suppl 1: i351-i358

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 35:7188-7196

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81:559-75

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature 468*:1053-1060

Renard C, Hart E, Sehra H, Beasley H, Coggill P, Howe K, Harrow J, Gilbert J, Sims S, Rogers J, Ando A, Shigenari A, Shiina T, Inoko H, Chardon P, Beck S (2006) The genomic sequence and analysis of the swine major histocompatibility complex. Genomics 88:96-110

Repbase: http://www.girinst.org/repbase/index.html

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: 205-211

Rothschild MF (2004) Porcine genomics delivers new tools and results: this little piggy did more than just go to market. Genet Res. 83:1-6

Saini SS, Hein WR, Kaushik A (1997) A single predominantly expressed polymorphic immunoglobulin VH gene family, related to mammalian group, I, clan, II, is identified in cattle. Mol Immunol. 34:641-51

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J (2010) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res.38:D5-16.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169-75.

Schook LB, Beever JE, Rogers J, Humphray S, Archibald A, Chardon P, Milan D, Rohrer G, Eversole K (2005) Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome. Comp Funct Genomics. 6251-5.

Schwartz JC, Lefranc MP, Murtaugh MP (2012) Organization, complexity and allelic diversity of the porcine (Sus scrofa domestica) immunoglobulin lambda locus. Immunogenetics. 64:399-407.

Schwartz JC, Lefranc MP, Murtaugh MP (2012) Evolution of the porcine (Sus scrofa domestica) immunoglobulin kappa locus through germline gene conversion. Immunogenetics. 64:303-311

Searle SM, Gilbert J, Iyer V, Clamp M. (2004) The otter annotation system. Genome Res. 14:963-970

Servin B, Faraut T, Iannuccelli N, Zelenika D, MilanD (2012) High-resolution autosomal map of the porcine genome using radiation-hybrid genotyping of the Illumina porcineSNP60 BeadChip: analysis and validation of the Pig genome assembly . Submitted

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Smit, AFA, Hubley, R & Green, P: RepeatMasker Open-3.0. 1996-2010. www.repeatmasker.org

Sperber GO, Airola T, Jern P, Blomberg J (2007) Automated recognition of retroviral sequences in genomic data--RetroTector. Nucleic Acids Res. 35:4964-76

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690

Strimmer K (2008) fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics. 24:1461-1462.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Project 1000 G, et al. (2010). Diversity of human copy number variation and multicopy genes. Science 39: 641 – 646.

Taliaferro JM, Alvarez N, Green RE, Blanchette M and Rio DC (2011) Evolution of a tissue-specific splicing factor network. Genes and Development 25:608-620

Tanaka-Matsuda M, Ando A, Rogel-Gaillard C, Chardon P, Uenishi H (2009) Difference in number of loci of swine leukocyte antigen classical class I genes among haplotypes. Genomics. 93:261-273

Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. (2006) Reconstructing genetic ancestry blocks in admixed individuals. American journal of human genetics 79:1-12

Talavera G and Castresana J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. System. Biol. 56: 564–577

Tortereau, F., Servin, B., Frantz, L., Megens, H.-J., Milan, D. et al. Sex specific differences in recombination rate in the pig are correlated with GC content. *Submitted* (2012).

Uenishi H, Hiraiwa H, Yamamoto R, Yasue H, Takagaki Y, Shiina T, Kikkawa E, Inoko H, Awata T. (2003) Genomic structure around joining segments and constant regions of swine T-cell receptor alpha/delta (TRA/TRD) locus. Immunology. 109:515-526.

Uenishi H, Eguchi-Ogawa T, Toki D, Morozumi T, Tanaka-Matsuda M, Shinkai H, Yamamoto R, Takagaki Y. (2009) Genomic sequence encoding diversity segments of the pig TCR delta chain gene demonstrates productivity of highly diversified repertoire. Mol Immunol. 46:1212-1221

Vandenbroeck K, Fiten P, Beuken E, Martens E, Janssen A, Van Damme J, Opdenakker G, Billiau A. (1993) Gene sequence, cDNA construction, expression in Escherichia coli and genetically approached purification of porcine interleukin-1 beta. Eur. J. Biochem. 217:45-52

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. 425:832-6.

vonHoldt, B. M., Pollinger, J. P., Earl, D. a, Knowles, J. C., Boyko, A. R., Parker, H., Geffen, E., et al. (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome research 21:1294-305

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blöcker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MC, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G, Fryc S, Guérin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J, Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras MF, Raudsepp T, Rocchi M, Røed KH, Ryder OA, Searle S, Skow L, Swinburne JE, Syvänen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC; Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team, Lander ES, Lindblad-Toh K. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326: 865-7.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.