

ARTICLE

Received 27 Jan 2014 | Accepted 12 Jun 2014 | Published 15 Jul 2014

DOI: 10.1038/ncomms5392

OPEN

# Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression

Mirte Bosse<sup>1</sup>, Hendrik-Jan Megens<sup>1</sup>, Laurent A.F. Frantz<sup>1</sup>, Ole Madsen<sup>1</sup>, Greger Larson<sup>2</sup>, Yogesh Paudel<sup>1</sup>, Naomi Duijvesteijn<sup>1,3</sup>, Barbara Harlizius<sup>3</sup>, Yanick Hagemeijer<sup>1</sup>, Richard P.M.A. Crooijmans<sup>1</sup> & Martien A.M. Groenen<sup>1</sup>

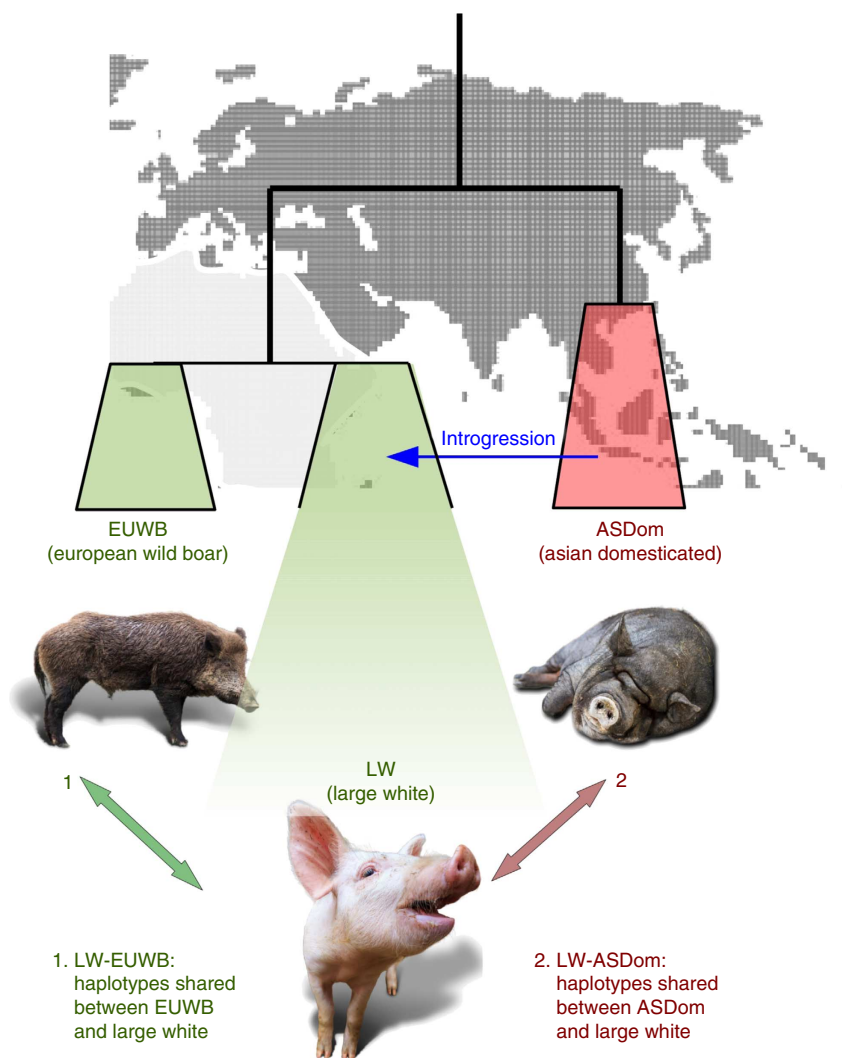
The independent domestication of local wild boar populations in Asia and Europe about 10,000 years ago led to distinct European and Asian pig breeds, each with very different phenotypic characteristics. During the Industrial Revolution, Chinese breeds were imported to Europe to improve commercial traits in European breeds. Here we demonstrate the presence of introgressed Asian haplotypes in European domestic pigs and selection signatures on some loci in these regions, using whole genome sequence data. The introgression signatures are widespread and the Asian haplotypes are rarely fixed. The Asian introgressed haplotypes are associated with regions harbouring genes involved in meat quality, development and fertility. We identify Asian-derived non-synonymous mutations in the *AHR* gene that associate with increased litter size in multiple European commercial lines. These findings demonstrate that increased fertility was an important breeding goal for early nineteenth century pig farmers, and that Asian variants of genes related to this trait were preferentially selected during the development of modern European pig breeds.

<sup>1</sup>Animal Breeding and Genomics Centre, Wageningen University, Wageningen 6708 WD, The Netherlands. <sup>2</sup>Durham Evolution and Ancient DNA, Department of Archaeology, Durham University, DH1 3LE Durham, UK. <sup>3</sup>TOPIGS Research Center IPG, 6640 AA Beuningen, The Netherlands. Correspondence and requests for materials should be addressed to M.B. (email: mirte.bosse@wur.nl).

Thousands of years of selection on numerous traits in domesticated species such as dogs<sup>1,2</sup>, cows<sup>3</sup> and chickens<sup>4</sup> has led to a wide range of distinct phenotypes that are not (or only rarely) observed in the wild. Improvement of breeding practices, which involves making crosses between different breeds and even (sub)species, has greatly improved productivity in intensive farming systems. Examples include the white seed colour in rice, shown to originate from a single mutation that swept through different subspecies following hybridization<sup>5</sup>, and the yellow skin allele that is fixed in the majority of modern western chicken breeds originating from admixture with the Grey junglefowl<sup>6</sup>. Human-mediated introgression of alleles is likely to have played a major role in the genomic architecture of many modern domestic species, including pigs.

The wild boar (*Sus scrofa*) originated ~4 million years ago in Southeast Asia and has since expanded its range over Eurasia, leading to the emergence of numerous geographically and genetically divergent populations<sup>7,8</sup>. The independent domestication of two of these populations in East Asia and western Eurasia led to distinct domesticated populations<sup>9–11</sup>. Hybridization and introgression between domestic pigs that originated from highly divergent wild populations has resulted in modern genomes that possess a mosaic of different

haplotypes<sup>12,13</sup>. Although some gene flow may have taken place before the nineteenth century, it was certainly extremely rare given the geographic distance between Asia and Europe, and the lack of any historical records describing the importation of Asian breeds into Europe before the nineteenth century (or vice-versa). Mitochondrial studies have suggested that the introgression was mostly female driven<sup>12</sup> and the introduction of Asian pigs into Europe at the onset of the Industrial Revolution in the late eighteenth and early nineteenth centuries has been particularly well documented<sup>14,15</sup>. In parallel with increasing intensification of farming at that time, British pig breeders sought to improve productivity of the local breeds, and did so, in part, by importing Chinese pigs. Chinese pigs were renowned for having great mothering characteristics, superior meat quality, strong resistance to diseases, better adaptation to living in sties and producing larger litters (>15 live born young). The selection for specific traits in European pig breeds has resulted in multiple selective sweeps in the genome of domesticates<sup>16</sup>. As European pig breeders deliberately introgressed Asian haplotypes into European local breeds, it is expected that the origin of haplotypes for which evidence of selection exists often stems from Asian introgression. Known examples include the *EDNRB*, *IGF2* and *KITLG* regions<sup>17–19</sup>, all of which the identified variants



**Figure 1 | Experimental setup for the introgression detection.** Arrows indicate the comparisons between groups that are used for the IBD detection. Individuals from the LW breed are used for all pairwise comparisons with individuals from two geographical and functional groups: EUWB and ASDom. The blue arrow indicates the human-mediated introgression from ASDom into LW.

have considerable effect on the phenotypes. Interestingly, the selection criteria shifted through time. When Asian lines were first introgressed into European pigs, fatness was selected for, although now leanness is preferentially selected. The Large White (LW) breed is one of most widely used breeds in commercial pig production and originated in the United Kingdom in the late nineteenth century. The breed is renowned for high growth rate, desirable carcass lean meat percentage and a desirable feed to body weight conversion ratio, but also for large litters, heavy milk production and great mothering characteristics. As LW is thought to have a hybrid origin, some of these traits are potentially the result of selected Chinese haplotypes. Because of deliberate introduction of Asian germplasm into European pigs and subsequent intensive artificial selection, Asian haplotypes in LW genomes are expected to be non-randomly distributed, but rather to be overrepresented in regions that contain genes or regulatory elements that are linked to traits relevant for production.

Here we test this hypothesis and identify specific gene variants bred into European pigs involved in key production traits. More specifically, we interrogated the genomes of LW pigs to reveal patterns of introgressed and selected haplotypes, to unravel the genomic consequences of human-mediated hybridization and artificial selection.

## Results

**Evidence of introgression.** We identified haplotypes in the LW pigs that were identical by descent (IBD) with individuals from both the original source of domestication, the European wild boars (EUWB), and the source of introgression, the Asian domestic group (ASDom, Fig. 1). Individuals from different locations in Europe were used to represent the source of domestication, whereas individuals from three different Asian breeds were used to represent the pool of putative introgressed haplotypes (see Supplementary Table 1 and Supplementary Methods for details). Average genetic differentiation ( $F_{st}$ , as defined by Weir and Cockerham<sup>20</sup>) between the LW and ASDom was 0.33 (s.d. 0.23, s.e. 0.0008), while the average  $F_{st}$  value between LW and EUWB was 0.16 (s.d. 0.17, s.e. 0.0006). These results show that the genomes of the LW pigs still share greater similarity with their EUWB ancestors than with ASDom. We used an other independent method to further verify the existence of gene flow between ASDom and LW after lineage divergence between Asian and European *Sus scrofa* ( $D$ -statistics<sup>21</sup>, see Methods). We computed this statistic for each possible trio between (LW, ASDom and EUWB), and (LW, EUWB, ASDom) so that a significantly negative  $D$  ( $Z < -4$ ) imply admixture between ASDom and LW. Our results demonstrate that all LW individuals possess roughly an equal degree of admixture with Asian pigs over their entire genomes, reflecting the human-mediated hybridization with Asian domestics in the late eighteenth and early nineteenth century ( $D = -0.083 \pm 0.015$ ,  $Z = -20$ ).

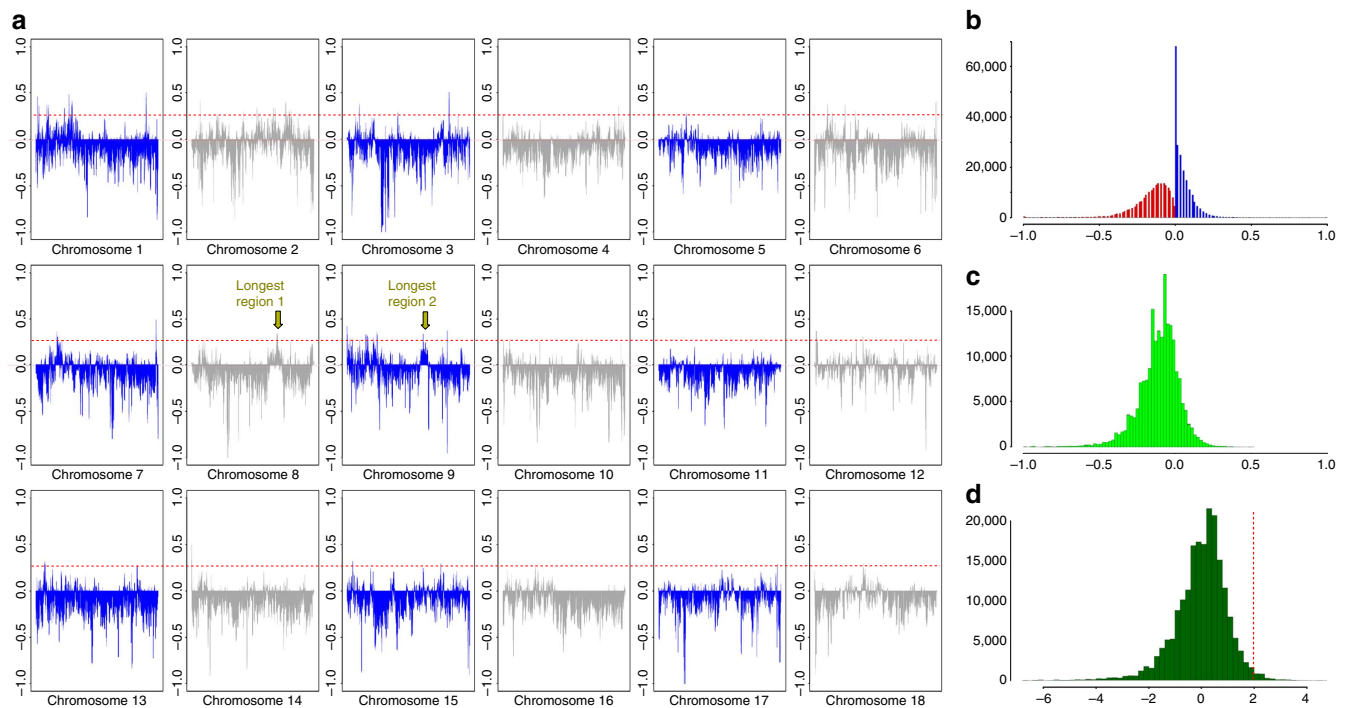
**Introgression mapping.** To infer whether a region was introgressed in multiple individuals, we calculated the frequency of all LW haplotypes that were of Asian or European origin for each bin of 10,000 bases in the genome. The relative fraction of Asian versus European haplotypes in the LW group is expressed as relative IBD (rIBD). Asian haplotype frequency in the LW population, for a given locus, ranged from 0.7 (where 1 indicates all haplotypes are ASDom and none are EUWB) to  $-1$  (all haplotypes are IBD with EUWB, Fig. 2a). The majority of the genome displays more similarity with the EUWB than with the Asian domesticated pigs. Despite this, every chromosome contained regions in which the signal for Asian ancestry was stronger than the European signature. A cutoff of two s.d. from the mean

in the  $Z$ -transformed rIBD distribution allowed us to identify regions, which spans  $\sim 1.3\%$  of the genome of LW pigs, that were likely to be of Asian origin (Fig. 2c,d and Supplementary Table 2).

This introgression pattern is probably the result of a combination of drift and selection. In contrast to dogs, where selection seems to have acted on a relatively small number of loci with large effects<sup>1,2</sup>, the introgression signal in the pigs is found at many loci, and the putatively selected Asian haplotype is rarely fixed. This pattern suggests that selection on Asian haplotypes, if present, mostly involved complex multi-genic traits or genes influencing traits selected in opposing directions. A high rIBD signal in our analysis refers to a region that contains predominantly introgressed haplotypes, but this does not imply that the introgressed haplotypes are identical or similar. Regions that contain more Asian haplotypes just by chance, but have not been under selection, could result in a high rIBD signal. We used an extended haplotype homozygosity test<sup>22</sup> to check for extended haplotypes in the LW population and compared the integrated haplotype score (iHS) signal with our rIBD values. This way, we can distinguish between regions that contain multiple Asian haplotypes and regions that contain one or few particular elongated Asian haplotypes. We inferred a significant correlation between rIBD and integrated haplotype score (iHS) in bins of 500 kb over the full genome (Supplementary Fig. 1a). To check whether the extended haplotype homozygosity in the LW pool was specific for the breed or observed in more European breeds, we contrasted the LW signal with a reference pool of other European commercial pigs (Supplementary Fig. 1b,c).

**Genome-wide patterns of introgression.** On a genome scale, many of the genes located within the regions where the LW pigs share more haplotypes with ASDom than with EUWB ( $\sigma ZrIBD \geq 2$ ), are associated with commercial traits such as meat quality (*DNMT3A*, *SAL1*, *ME1*, *IGF2BP1*), fertility (*PGRMC2*, *KIF18A*, *CDK20*, *AHR*) and development (*NRG1*, *AHR*, Supplementary Table 2), although no significant enriched Gene Ontology (GO) term was found. Gene-dense regions on chromosome 1 and 2 display a high rate of alternating ASDom and EUWB haplotypes. For instance, the regions containing the *CDK20* and *SAL1* genes, which both have been associated with reproduction traits<sup>23,24</sup>, are only 10–20 kb long. These short tracts of shared haplotypes either indicate a high recombination frequency (corroborated by the recombination map for pig<sup>25</sup>), a more temporally distant hybridization episode and/or favourable European haplotypes surrounding these genes that could lead to positive selection on recombinant haplotypes. The recombination landscape in *Sus scrofa* is known to be highly heterogeneous, and this probably results in an unequal distribution of haplotype length<sup>26</sup>. Longer Asian haplotypes will be found in regions of low recombination and therefore the introgression signal is easier to identify in regions with a low recombination rate.

**Longest regions of introgression.** Chromosomes 8 and 9 contain the largest consecutive regions of inferred introgression in the LW genomes (defined as regions where rIBD > 0). To check whether the extended haplotype homozygosity in the LW pool was specific for the breed or observed in more European breeds, we contrasted the LW signal with a reference pool of other European commercial pigs with the  $R_{sb}$  statistic<sup>27</sup> (Supplementary Fig. 1b,c). This analysis demonstrates that the region of introgression on chromosome 8 contains a stronger extended haplotype homozygosity (EHH) signal in the reference panel, and that the region on chromosome 9 contains a particularly strong signal in the LW population. We used two independent methods,  $D$ -statistics and  $F_{st}$ , to support the

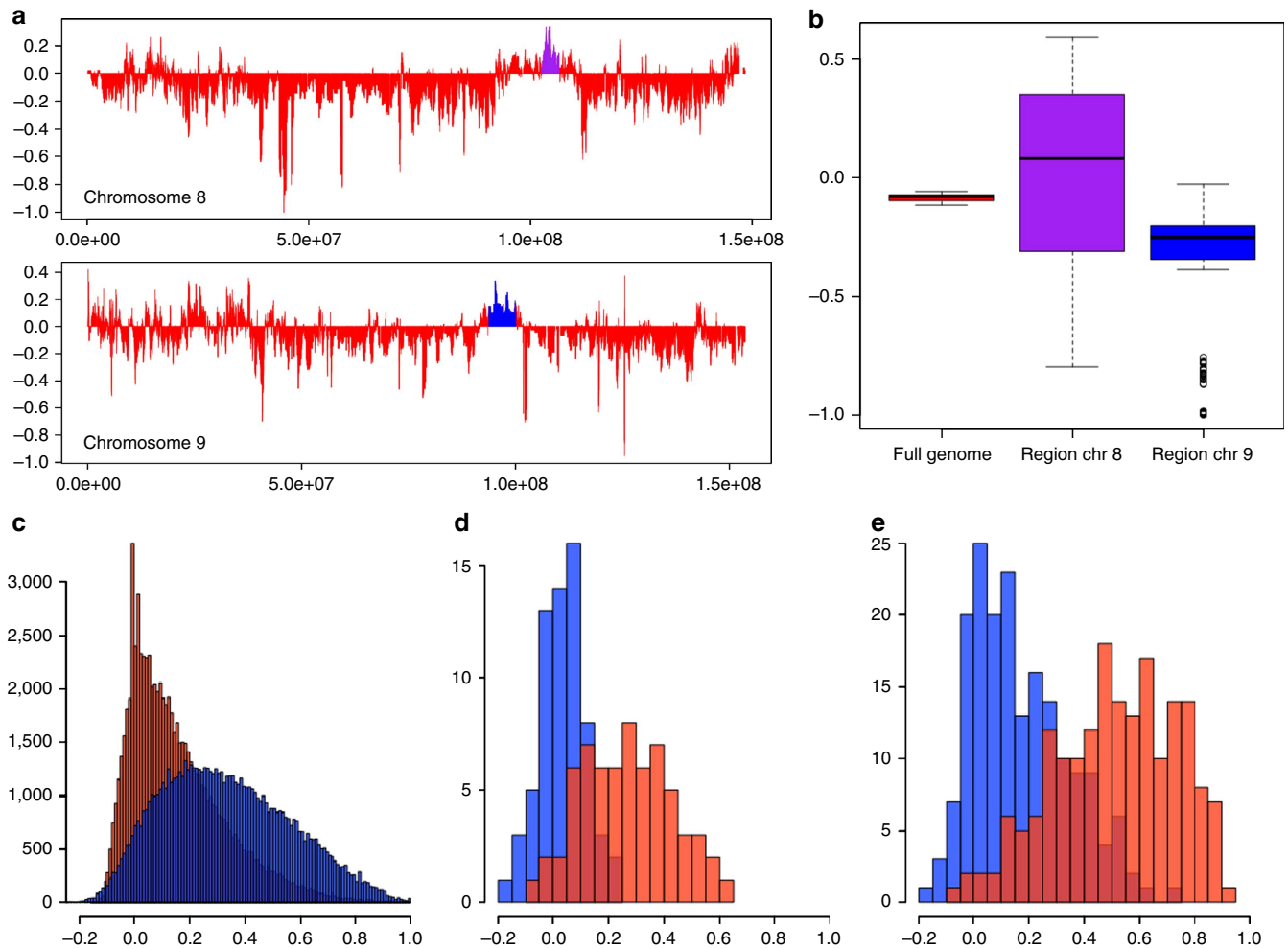


**Figure 2 | Distribution of regions in the genome where the LW contain introgressed haplotypes from ASDom.** (a) The x axis shows the full length of all chromosomes, and the y axis represents the relative frequency of LW haplotypes IBD with ASDom or EUWB, ranging from 1 (all haplotypes are IBD with ASDom, and none with EUWB) to  $-1$ . The two longest regions of consecutive introgression are indicated with arrows. (b) Distribution of the relative proportion of IBD haplotypes in LW and the EUWB (green, IBDEUWB, 0 to  $-1$ ) or ASDom (blue, IBDASDom, 0 to 1) in bins of 10 Kbp. (c) Distribution of the rIBD scores for the LW haplotypes ( $rIBD = IBDEUWB - IBDASDom$ ). (d) Z-transformed distribution of rIBD.

detected introgression in these regions in the LW (Fig. 3a–e). To show that divergence between LW and ASDom was reduced in the introgressed regions, we calculated  $F_{st}$  for these regions separately. The  $F_{st}$  between ASDom and the LW was lower in both introgressed regions than between EUWB and LW (Fig. 3c–e), thereby supporting the signal of Asian introgression (high rIBD). The  $D$ -statistics for the regions on chromosome 9 was lower than the genome-wide average, which corroborated our rIBD analysis (Fig. 3b). The region on chromosome 8 shows a wide distribution, indicating that some LW haplotypes contain the Asian signature, while others do not. Inconsistent clustering of European haplotypes within an Asian clade at this locus supports this hypothesis (Supplementary Fig. 2). Curiously, the  $\sim 4$  Mb sequence shows a clear signal of introgression, although a large part of the region is devoid of annotated genes. As this part of the genome has a relatively low recombination frequency<sup>25</sup>, the region may extend considerably beyond the position of the actual favourable allele that has been selected for, due to genetic hitchhiking and the short time since introgression. Alternatively, drift could have resulted in the presence of Asian haplotypes in this region. The *PGRMC2* gene, coding for the progesterone receptor, lies within the highest peak of Asian haplotypes in that region. Progesterone is an important hormone involved in female reproduction and maternal behaviour<sup>28</sup>, traits that Asian pigs have been selected for extensively. Therefore, the Asian haplotype containing the *PGRMC2* gene could be associated with higher reproductive success in LW pigs and may have been subjected to selection pressure as a result. The Rsb signal suggest that in other European breeds, the proportion of Asian haplotypes is even higher for this locus (Supplementary Fig. 1b,c). We used genotype data from the Illumina Porcine 60K iSelect beadchip<sup>29</sup> for an additional 5,143 pigs from three European commercial lines to screen allele frequencies in this region. Two genetic lines have been selected for reproductive traits since the establishment of the

lines (A and B), and one line for finishing traits (C). The SNP alleles in this 4-Mb region show a clear difference between the two reproduction-associated lines and the growth-associated line (Supplementary Fig. 3). These findings could indicate that the Asian haplotypes in this region are associated with fertility, but further analyses are needed to support this hypothesis.

**Introgression at the *AHR* locus.** The 6.8-Mb region on chromosome 9 has a large proportion of LW haplotypes that are nearly identical to two haplotypes found in the Asian Jianqubai breed (Supplementary Fig. 2). Among the genes in this region of putative Asian introgression are multiple members of the *TWIST* gene family, transcription factors known to be involved in a variety of processes, including embryonic development<sup>30</sup>. The highest introgression peak in this region contains the *AHR* gene that has previously been associated with female reproductivity<sup>31</sup>. Originally, the *AHR* gene has been identified as mediator of xenobiotic-induced toxicity<sup>32,33</sup>. *AHR* has been shown not only to be involved in the response to toxicity but also to be associated with fertility in mammals<sup>34,35</sup>. The *AHR* gene seems to play an important role in the female reproductive system at multiple levels<sup>31</sup>. In pigs, expression of *AHR* during the oestrous cycle and putative involvement of *AHR* in the regulation of reproduction has been observed<sup>36,37</sup>. Furthermore, polymorphisms in human are known to occur predominantly in exon 10, which contains an important transactivation domain<sup>38</sup>. We screened this gene in pigs for non-synonymous mutations and identified four non-synonymous mutations in exon 11 of the *AHR* gene, which corresponds to exon 10 of the human *AHR* gene. The variants of Asian origin were all in strong linkage disequilibrium. As the *AHR* is a strong candidate gene for which Asian variants have been selected since introgression during the Industrial Revolution, we examined this gene further (see Methods).



**Figure 3 | Levels of differentiation between LW and ASDom or EUWB in regions of introgression.** (a) Relative introgression fraction (rIBD) over the full length of chromosome 8 and 9. The longest regions of introgression are indicated with purple and blue. (b) Boxplot of  $D$ -statistics for the full genome (red) and the two longest regions of introgression as indicated in a on chr8 (purple) and chr9 (blue). The minimum, first quartile, median, third quartile and maximum are indicated with the box and whiskers with outliers  $>1.5 \times \text{IQR}$ .  $D$ -statistics are computed for each possible trio with LW = P1, ASDom = P2 and EUWB = P3, with the Sumatran *S. scrofa* as outgroup (O) resulting in 378 trios. (c–e) Distribution of  $F_{st}$  between LW-ASDom (blue) and LW-EUWB (red) in bins of 10 Kbp. The left histogram shows the  $F_{st}$  distributions based on the full genome (c), and the other two show the  $F_{st}$  distribution for the regions of introgression on chr8 (d) and chr9 (e).

The iHS signal within the LW population is strongest for the AHR locus. The fact that the Rsb signal is also strong in the LW compared with other European breeds, indicates that the Asian haplotypes at the AHR locus were never fixed in the ancestral population that led to current commercial breeds, because the frequency and iHS signal differs between breeds (Supplementary Fig. 1b,c). We used an additional method to screen for selection, nSL, that has recently become available and should be robust to variation in recombination rate<sup>39</sup>. We averaged the rIBD, the nSL statistic and the corrected iHS  $P$ -value over bins of 500 kb for chromosome 9 (see Supplementary Fig. 4a–c). All three statistics contain a high signal at the previously identified region on chromosome 9 that also contains the AHR gene. We show that the haplotype containing the ancestral allele at the AHR locus is more persistent than haplotypes containing the derived allele at this locus (Supplementary Fig. 4d–f). A survey of wild boar populations and four other *Sus* species for these loci revealed that the ancestral haplotype is homozygous in all closely related *Sus* species, and at high frequency in Asian domesticated pigs and European breeds. However, only derived haplotypes were found in European wild boar and the ancestral type was present only at

a low frequency in Asian wild boars. This suggests a history of selection for the ancestral state in domestics, after the derived state reached high frequency in the wild populations.

**Effect of Asian haplotypes at the AHR locus.** To examine the effect of the amino acid changes in the AHR protein on reproductive success, we used genotype data from the Illumina Porcine 60K iSelect beadchip<sup>29</sup> for the same 5,143 pigs from three European commercial purebred lines for which estimated breeding values (EBV) for the total number of piglets born (TNB) were available. We extracted genotypes for those markers that fell within the region of high introgression (rIBD  $> 2$ ). The AHR gene was the only annotated gene in this selected part of the genome. Either Asian or European heritage was assigned to the 60K haplotypes for these 5,143 commercial pigs at the AHR locus, based on our re-sequenced individuals and confirmation in the lab (see Methods). Haplotypes containing the Asian AHR variants had a significant effect on the EBV for total number born over all three lines (EBV–TNB 0.162, s.e. 0.04,  $P < 0.0001$ , Table 1). Although the EBV–TNB can be rather different between different

**Table 1 | Estimated breeding values for total number born per line for haplotypes containing Asian or European alleles at the *AHR* locus.**

Line	N	Linear model				Animal model					
		As-EBV	Eu-EBV	Estimate	s.e.	t-value	Significance	Effect	Se effect	F-value	P-value
All	2,586	0.038	-0.115	-0.163	0.039	-4.221	2.54e-05***	-0.07	0.03	5.65	0.018*
Line C	965	-0.091	-0.178	-0.089	0.040	-2.191	0.0287*	-0.04	0.09	0.19	0.66
Line B	568	0.242	0.113	-0.128	0.088	-1.461	0.144	-0.02	0.02	0.91	0.342
Line A	1,053	0.039	-0.167	-0.227	0.068	-3.329	0.000901***	-0.15	0.06	6.5	0.012*

EBV, estimated breeding value; TNB, total number of piglets born.

Two models were used to estimate the effect of the Asian allele: a linear model and an animal model. As-EBV is the EBV for haplotypes containing the Asian alleles and Eu-EBV is the EBV for European haplotypes under the linear model. 'Estimate' is the estimated effect of the origin of the haplotypes on TNB for the linear model and 'Effect' refers to the effect of the origin of haplotypes under the animal model. The effect was calculated for two reproduction-associated lines (line A, N=1,053 and Line B, N=568) and one growth-associated line (Line C, N=965). N is number of haplotypes in the line. Significant codes: 0.001\*\*\* 0.01\*\* 0.05\* 0.1.

lines, within-line effects of the Asian allele were also estimated. A difference of 0.1 in EBV equates to a difference of 0.1 piglets born. As total number of piglets born is a complex multi-locus trait, an increase of 0.16 piglets born (across all three breeds) is substantial in the current breeding industry. If the costs of maintaining a sow on a farm are spread over a larger number of piglets being weaned from that sow, the marginal cost reduction of producing a finishing pig is just over 3 euros per extra piglet<sup>40</sup>, that is, 2% of that total. Even though we cannot rule out that the effect could be due to some extended haplotypes covering other genes in the region, these details in combination with known literature contribute to *AHR* as the strongest candidate gene.

Interestingly, the *AHR* locus is an example of selection on the ancestral state that was either lost or never present in the European wild population. Re-introduction of the variants by introgression of Asian haplotypes and the positive effect of these alleles on litter size contributed to the high frequency of Asian haplotypes at the *AHR* locus in the current population. As the *AHR* gene seems to be involved in multiple life history traits, it could very well be that some long-term balancing selection acted on the alleles. The *AHR* gene is involved in multiple traits and during ever-changing adaptation to, for example, different environments, some alleles might be more desirable than others under different circumstances<sup>41</sup>. The significant association between the Chinese haplotypes and an increased EBV for total number born, which in our opinion is a strong independent indication for selection, in combination with the selection sweep results provide convincing evidence for the *AHR* locus to have been under selection after introgression. Similar examples of selection on Asian haplotypes in European pigs exist in literature, such as the signals of selection associated with coat colour<sup>42</sup>, ear morphology<sup>17</sup> and increased lean content<sup>19</sup>.

The evidence presented here demonstrates how crossing of divergent populations may shape the variation on a genome-wide scale in populations. The introduction of Asian haplotypes into European breeds in the late eighteenth and early nineteenth centuries<sup>8,9</sup> and consecutive selection for desired traits in these breeds thereby provides a robust, historically documented model system for these instances. We identified numerous genomic regions where Asian haplotypes were introgressed into a larger European background, including the *AHR* locus. The *AHR* gene has been known to be involved in reproduction<sup>34-37</sup>, and our study corroborates that earlier report by demonstrating a significant increase in litter size in European commercial pigs that possess the Asian haplotype. Our findings provide a unique insight into the genomic haplotype patterns resulting from breeding practices from first domestication until the intensive breeding industry we know today. The observed introgression pattern is a combination of drift and selection, and detailed analyses such as those demonstrated for the *AHR* locus will shed

more light onto the importance of other introgressed Asian haplotypes on signatures of selection in modern pig breeds.

## Methods

**Sample collection and DNA preparation.** Blood samples were collected from a total of 70 individual, wild and domesticated *S. scrofa*. Among these individuals were 2 wild boars from Sumatra, 8 Asian wild boars from China and Japan, 18 EUWB from the Netherlands, France, Switzerland, Greece and Italy, 13 Asian domesticated pigs from the Meishan, Jianquhai and Xiang breeds and 29 European domesticated pigs from the Duroc, Hampshire, Pietrain, Landrace and LW breeds. DNA was extracted from the blood samples with the use of the QIAamp DNA blood spin kit (Qiagen Sciences) and checked for quality and quantity on the Qubit 2.0 fluorometer (Invitrogen). Library construction for the re-sequencing was performed with 1–3 µg of genomic DNA according to the Illumina library prepping protocols (Illumina Inc.) and the insert size varied from 300 to 500 bp. Sequencing was performed on 1–3 µg of genomic DNA with the 100 paired-end sequencing kit for all samples. Single nucleotide polymorphism (SNP) genotyping was performed on the Illumina Porcine 60K iSelect Beadchip<sup>26</sup>. DNA from all individuals was diluted to 100 ng µl and genotyped according the IlluminaHD iSelect protocol. Data was analysed using Genome Studio software (Illumina Inc.).

**Alignment and variant calling.** All individuals were re-sequenced with the Illumina paired-end sequencing technology (Illumina Inc.) to ~10 × depth of coverage (details in Supplementary Table 1). The read pairs were trimmed to have a minimum phred quality >20 over three consecutive bases, while each mate should have a minimal size of 45 bp after trimming. Alignment was performed with Mosaik Aligner (V. 1.1.0017) with the unique alignment option to the Porcine reference genome build 10.2. Variants were called using Samtools mpileup 0.1.12a (r862). The alternative allele should be covered at least two times to call an SNP and INDELS were excluded. VCFtools was used for filtering for a genotype quality of >20, a minimum read depth of 7 × and a maximum read depth of twice the average read depth. For the list of all those sites that were heterozygous or non-reference within at least one individual, genotyping was performed with Samtools mpileup for all 70 individuals to create a matrix containing an unbiased representation of the variation present in the samples. Homozygous reference alleles were also included in the matrix at this stage, and only those sites that were covered ≥4 × in all individuals were included, resulting in 2,377,607 markers.

**Pairwise IBD detection.** The matrix of 70 individuals genotyped for 2,377,607 positions in the genome served as input for the IBD detection pipeline. All individuals were phased with the fastPhase function in Beagle version 3.3.2. Pairwise shared haplotypes were extracted with the Beagle fastIBD function as described by Browning and Browning<sup>43</sup>. Phasing and IBD detection was executed ten times independently and identified IBD tracts were merged from all ten runs, as suggested by the authors. Partially overlapping runs were extracted and the IBD runs with the highest probability scores were added to the pool of IBD tracts. The ten cycles of IBD detection were run with different thresholds for assigning IBD to the haplotypes of two individuals. The numbers varied from zero detected pairwise IBD tracts to complete IBD genomes. We empirically determined that the middle range of thresholds resulted in pairwise IBD tracts that remained stable in terms of relative number and length of detected IBD tracts between members of different pig groups. To this end, the final threshold that was used for IBD detection (5.0<sup>-6</sup>) was elevated compared with that of the original paper, to allow extracting similar, but not necessarily identical, haplotypes between individuals. As the focus of this analysis was to identify regions containing haplotypes that are more similar between distantly related individuals than expected based on their inheritance, and the frequencies of similar haplotypes are leveled out, this threshold was thought to fit the data best.

To estimate the frequency of shared haplotypes in different regions of the genome, the genome was divided into bins of 10,000 bp, and the number of recorded IBD tracts between the LW pigs and the two different pig groups (ASDom and EUWB) was computed per bin. As the total number of pairwise comparisons differed between the groups, these numbers were normalized, ranging from 0 (no IBD tract detected) to 1 (all individuals IBD with all individuals within the group). Relative IBD between the LW and the two competing pig groups ASDom and EUWB was then calculated by extracting per bin the normalized IBD with ASDom from the normalized IBD with EUWB.

Normalized IBD for one pig group:  $rIBD = cIBD/tIBD$  (where  $cIBD$  = count of all haplotypes IBD between LW and one pig group and  $tIBD$  = total pairwise comparisons between LW and one pig group).

$rIBD$  between two pig groups:  $rIBD = nIBDASDom - nIBDEUWB$ .

The distribution of  $rIBD$  for the comparison between LW-ASBr and LW-EUWB IBD haplotypes resembled a normal distribution and therefore was Z-transformed as follows:  $ZrIBD = (rIBD - \mu) / \sigma rIBD$ . Therefore,  $ZrIBD$  represents the number of s.d. that  $rIBD$  deviates from the mean  $rIBD$ . The threshold for extreme IBD with the breeds from the other continent compared with the wild boars from the same continent was set to 2 s.d. from the mean in the far right tail of the distribution.

**GO-enrichment analysis.** All annotated genes in build 10.2 (Ensembl release 67) from the *S. scrofa* reference genome were extracted. GO-enrichment analysis was performed for genes in the top 1.3% ( $\sigma > 2$  for  $ZrIBD$ ) of regions with an over-representation of ASDom haplotypes in LW pigs. The Cytoscape v.2.8.3 plugin BinGO v2.44 (ref. 44) was used to identify over-represented biological process-related GO terms. The human one-to-one orthologues (Ensembl db) for all pig genes were used for the analysis, as human genes are annotated more comprehensively. Significance levels were adjusted based on the Benjamini and Hochberg correction for multiple comparisons.

**Fst analysis.** To measure genetic differentiation the individuals from the Matrix were assigned to one of the following pig groups (if applicable): LW breed (LW), Asian domesticated pigs (ASDom) and European wild boar (EUWB). Pairwise Fst as described by Weir and Cockerham<sup>20</sup> between the LW breed and the other two groups were computed with Genepop 4.2 in bins of 10 kb over the full length of the genome<sup>45</sup>. Relative Fst ( $rFst$ ) and Z transformation of  $rFst$  ( $ZrFst$ ) were computed similarly as  $rIBD$  and  $ZrIBD$ . Correlations between  $ZrFst$  and  $ZrIBD$  values were calculated with Pearson's product moment correlation in R.

**Admixture fraction.** To prove the existence of admixture between LW and MS in our potentially introgressed regions, but also genome-wide, we computed *D*-statistics<sup>21</sup> as implemented in qpDstats from the ADMIXTOOL software package<sup>46</sup>. In short, the *D*-statistics provide a robust test for admixture by challenging the strictly bifurcating nature of a phylogenetic tree. For a triplet of taxa P1, P2 and P3, and an outgroup O, that follows the phylogeny ((P1,P2),P3),O, one can compute the number of sites where P1 and P3 (BABA) or P2 and P3 (ABBA) share the derived state (B; assuming ancestral state, A, in the outgroup). Under a null hypothesis of no gene flow or no sub-structure (strict bifurcation), the count of ABBA and BABA should not be significantly different. Alternatively, a significant excess of either ABBA or BABA site pattern provide a conclusive proof of gene flow or sub-structure. However, as sub-structure is very unlikely to affect our analysis of domestic pigs from Asia and Europe (because of independent domestication), we can conclude that significant *D* implies gene flow. We computed *D*-statistics between LW (P1), European wild boars (P2) and Asian domestics (P3), for every possible combination of samples, using the sequence of wild boar from Sumatra as an outgroup (O). For each possible combination, we first computed a genome-wide value. Significance level was computed using a standard block jackknife, with blocks of 1 cM (assuming 1 Mb = 1 cM). We also computed *D*-statistics in potentially admixed regions separately.

**Haplotype association test.** To examine the putative effect of the amino acid change, we used genotype data from an additional 5,143 individuals from 3 different commercial purebred lines, genotyped on the Illumina Porcine 60K iSelect Beadchip<sup>29</sup> for 21 markers surrounding the mutation. Line A and B are dam lines selected for reproductive traits and line C is a sire line selected for finishing traits. In all three lines, TNB was routinely recorded on sows. The EBV of the genotyped animals were obtained via routine genetic evaluation using MIXBLUP in a multitrait model<sup>47</sup>. The model for obtaining the EBV of TNB included fixed effects (herd-year-season, insemination number, parity, cross-fostering (Y/N), interval weaning (class)) and random effects for service sire, permanent sow and animal. Reliabilities per animal were extracted from the genetic evaluation and were based on the methodology of Tier and Meyer<sup>48</sup>, and animals with a reliability < 0.15 were excluded from the analyses. We genotyped 64 individuals for the mutation by Sanger sequencing with the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) on an ABI3730 sequencer. The following primers were used: AHRgene\_Forward 5'-AGCAGCAGCAACAACACTGTGT-3' and AHRgene\_Reverse 5'-GACACAGCTCCACCATAGCA-3'. The haplotypes, based on the 21 markers that were associated with the alleles at the site of mutation, were reconstructed.

Next, all 5,143 individuals were phased with Beagle, and the G or T allele at the site of the mutation was assigned to these phased haplotypes when possible. We identified 19 unique haplotypes in the data set for which the allele at the site of mutation could be verified. The following linear model was fit to the data using R to test for a significant effect of the allele on EBV for TNB:

$lm(\text{formula} = \text{EBV\_TNB} \sim \text{LINE} + \text{Origin}, \text{weights} = \text{REL\_TNB})$ . EBV\_TNB is the EBV for total number born and these values are weighted for the reliability for this trait (REL\_TNB). Genetic line was included to account for differences between selection strategies between the three genetic lines A, B and C.

Finally, an animal model was fit (to account for family relatedness) to the data using ASReml<sup>49</sup> to test for a significant effect of the allele on EBV for TNB:

$y = Xb + Za + e$ , where  $b$  is a vector of the fixed effect for line and origin,  $a$  is the vector of the random animal genetic effect. The term  $e$  is a vector of the random residual effects assumed to be normally distributed, but weighted by the reliability of the EBV.

**Mutation characteristics.** The functional annotation of the genomic variants in the high IBD regions ( $\sigma rIBD > 2$ ) was determined using Annovar<sup>50</sup>. The nature of the non-synonymous mutation in the gene was obtained from the webtool polyPhen 2 (ref. 51) by using the human orthologue *AHR* for the pig gene ENSSSCG00000030484.

**Phylogenetics.** We performed our primary phylogenetic analysis using MrBayes 3.2 and our matrix of variable sites<sup>52</sup>. To estimate correct branch length solely from variable sites, we used the Mk model implemented in MrBayes, which provides a likelihood framework for data sets that contains only variable characters. We recorded SNPs in four potential states (0–3). Rate of evolutions, for each SNP, were drawn from a  $\gamma$ -distribution. We ran two independent runs of four Markov chain Monte Carlo (MCMC) chains with two million samples. We repeated this analysis solely based on SNPs found in the two introgressed regions on chromosome 9 and 8 separately.

**Extended haplotype homozygosity tests.** The identification of extended haplotype homozygosity was tested on 56 LW individuals that were genotyped on the Illumina Porcine 60K iSelect Beadchip<sup>29</sup>. First, a genome-wide scan for iHS within line was performed with the R package rehh<sup>22,53</sup>. Significance levels within line were averaged for 500 kb and correlation with the  $rIBD$  values for the same bins of 500 kb was tested with the cor.test R. Second, a reference panel of 100 individuals belonging to the Landrace, Pietrain and another LW breed was used to polarize the iHS signal in the original 56 LW individuals with the ies2rsb function in rehh<sup>27,53</sup>. To check the signal on chromosome 9, we also performed the recently developed nSL test that uses a slightly different approach to screen for extended haplotype homozygosity than the original iHS test and is robust to variation in recombination frequency<sup>39</sup>. The bifurcation diagram option in rehh was used to visualize the linkage disequilibrium from a focal SNP on the Beadchip that was closest to the *AHR* gene.

## References

- Boyko, A. R. *et al.* A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol.* **8**, e1000451 (2010).
- vonHoldt, B. M. *et al.* Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**, 898–902 (2010).
- Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700–707 (2002).
- Miao, Y. -W. *et al.* Chicken domestication: an updated perspective based on mitochondrial genomes. *Heredity* **110**, 277–282 (2012).
- Sweeney, M. T. *et al.* Global dissemination of a single mutation conferring white pericarp in rice. *PLoS Genet.* **3**, e133 (2007).
- Eriksson, J. *et al.* Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. *PLoS Genet.* **4**, e1000010 (2008).
- Frantz, L. A. F. *et al.* Genomic sequencing provides fine scale inference of evolutionary history. *Genome Biol.* **2013**, R107 (2013).
- Meijaard, E., d'Huart, J.P. & Oliver, W.L.R. in: *Handbook of the Mammals of the World Vol 2* (eds Wilson, D.E. & Mittermeier, R.A.) 248–291 (Lynx Edicions, Barcelona, Spain, pp. 2011).
- Larson, G. *et al.* Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* **307**, 1618–1621 (2005).
- Megens, H. J. *et al.* 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**, 103–128 (2008).
- Groenen, M. A. M. *et al.* Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393–398 (2012).
- Giuffra, E. *et al.* The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* **154**, 1785–1791 (2000).
- Goedbloed, D. *et al.* Genome-wide SNP analysis reveals recent genetic introgression from domestic pigs into Northwest European wild boar populations. *Mol. Ecol.* **22**, 856–866 (2013).

14. White, S. From globalized pig breeds to capitalist pigs: a study in animals cultures and evolutionary history. *Environ. Hist.* **16**, 94–120 (2011).
15. Jones, G. F. in *The Genetics of the Pig* (eds Ruvinsky, A. & Rothschild, M. F.) 17–50 (CAB International, Oxon, UK, 1998).
16. Rubin, C. -J. et al. Strong signatures of selection in the domestic pig genome. *Proc. Natl Acad. Sci. USA* **109**, 19529–19536 (2012).
17. Wilkinson, S. et al. Signatures of diversifying selection in European pig breeds. *PLoS Genet.* **9**, e1003453 (2013).
18. Okumura, N., Matsumoto, T., Hamasima, N. & Awata, T. Single nucleotide polymorphisms of the KIT and KITLG genes in pigs. *Anim. Sci. J.* **79**, 303–313 (2008).
19. Ojeda, A. et al. Selection in the making: a worldwide survey of haplotypic diversity around a causative mutation in porcine IGF2. *Genetics* **178**, 1639–1652 (2008).
20. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358 (1984).
21. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
22. Sabeti, P. C. et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
23. Liu, Y. G. & Xia, X. H. Molecular characterization, polymorphism and association with reproductive traits of porcine CDK20 gene. *J. Anim. Feed Sci.* **20**, 566–574 (2011).
24. Seo, H., Kim, M., Choi, Y. & Ka, H. Salivary lipocalin is uniquely expressed in the uterine endometrial glands at the time of conceptus implantation and induced by interleukin-1beta in pigs. *Biol. Reprod.* **84**, 279–287 (2011).
25. Tortoreau, F. et al. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics* **13**, 586 (2012).
26. Bosse, M. et al. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS Genet.* **8**, e1003100 (2012).
27. Tang, K., Thornton, K. R. & Stoneking, M. A new approach for using genome scans to detect recent positive selection in the human genome. *Plos Biol.* **7**, e171 (2007).
28. Chen, C. et al. Cloning, mapping and molecular characterization of porcine progesterone receptor membrane component 2 (*PGRMC2*) gene. *Genet. Mol. Biol.* **33**, 471–474 (2010).
29. Ramos, A. M. et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE* **4**, e6524 (2009).
30. Gitelman, I. Evolution of the vertebrate twist family and synfunctionalization: a mechanism for differential gene loss through merging of expression domains. *Mol. Biol. Evol.* **24**, 1912–1925 (2007).
31. Hernández-Ochoa, I., Karman, B. N. & Flaws, J. A. The role of the aryl hydrocarbon receptor in the female reproductive system. *Biochem. Pharmacol.* **11**, 547–559 (2009).
32. Fujii-Kuriyama, Y. & Kawajiri, K. Molecular mechanisms of the physiological functions of the aryl hydrocarbon (dioxin) receptor, a multifunctional regulator that senses and responds to environmental stimuli. *Proc. Jpn Acad. B* **86**, 40–53 (2010).
33. Denison, M. S., Soshilov, A. A., He, G., DeGroot, D. E. & Zhao, B. Exactly the same but different: promiscuity and diversity in the molecular mechanisms of action of the aryl hydrocarbon (dioxin) receptor. *Toxicol. Sci.* **124**, 1–22 (2011).
34. Baba, T. et al. Intrinsic function of the aryl hydrocarbon (dioxin) receptor as a key factor in female reproduction. *Mol. Cell Biol.* **25**, 10040–10051 (2005).
35. Pocar, P., Fischer, B., Klonisch, T. & Hombach-Klonisch, S. Molecular interactions of the aryl hydrocarbon receptor and its biological and toxicological relevance for reproduction. *Reproduction* **129**, 379–389 (2005).
36. Jablonska, O. et al. The expression of the aryl hydrocarbon receptor in reproductive and neuroendocrine tissues during the estrous cycle in the pig. *Anim. Reprod. Sci.* **126**, 221–228 (2011).
37. Onteru, S. K. et al. A whole-genome association study for pig reproductive traits. *Anim. Genet.* **43**, 18–26 (2012).
38. Harper, P. A., Wong, J. M. Y., Lam, M. S. M. & Okey, A. B. Polymorphisms in the human AH receptor. *Chem. Biol. Interact.* **141**, 161–187 (2002).
39. Ferrer-Admetlla, A., Liang, M., Korneliusen, T. S. & Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.* **31**, 1275–1291 (2014).
40. Hanenberg, E. H. A. T., Mathur, P. K. & Knol, E. F. Marginal economic values for pig production in different countries. in: 9th World congress on genetics applied to livestock production Leipzig, Germany (2010).
41. Connallon, T. & Clark, A. G. Antagonistic versus non-antagonistic models of balancing selection: characterizing the relative timescales and hitchhiking effects of partial selective sweeps. *Evolution* **67**, 908–917 (2013).
42. Kijas, J. M. H. et al. Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. *Genetics* **150**, 1177–1185 (1998).
43. Browning, B. L. & Browning, S. R. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **88**, 173–182 (2011).
44. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **21**, 3448–3449 (2005).
45. Rousset, F. Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* **8**, 103–106 (2008).
46. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
47. Mulder, H. A., Lidauer, M., Strandén, I., Mäntysaari, E. A., Pool, M.H. & Veerkamp, R.F. *MiXBLUP Manual* (Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, The Netherlands, 2012).
48. Tier, B. & Meyer, K. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. *J. Anim. Breed Genet.* **121**, 77–89 (2004).
49. Gilmour, A. R. et al. ASReml user guide release 3.0.VSN International Ltd, Hemel Hempstead, UK, 2009).
50. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
51. Adzhubei, I. A., Schmidt, S. & Peshkin, L. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
52. Ronquist, F. & Huelsenbeck, J. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
53. Gautier, M. & Vitalis, R. rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176–1177 (2012).

## Acknowledgements

DNA samples were provided by Dr Ning Li, China Agricultural University, China; Dr Alain Ducos, UMR INRA-ENVIT, France; Sem Genini, Parco tecnologico Padano, Italy; Dr Gono Semiadi, Puslit Biologi, Indonesia; Dr Naohiko Okumura, Staff Institute 446-1 Ippaizuka, Japan; Dr Alan Archibald, Roslin Institute and the Royal (Dick) School of Veterinary Studies, University of Edinburgh, Scotland; TOPIGS Research Center IPG BV, The Netherlands; Dr Oliver Ryder, San Diego Zoo, USA; Cheryl L. Morri, Ph.D., Omaha's Henry Doorly Zoo, USA. We thank Bert Dibbitts for the genotype validation assay in the lab, Miguel Perez Enciso for valuable comments on the manuscript and B. van de Water for graphics production. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013)/ERC Grant agreement number 249894.

## Author contributions

The experiment was designed by M.B., H.-J.M., O.M. and M.A.M.G. Sampling and data preparation was done by R.P.M.C., H.-J.M., G.L., B.H. and N.D. Bioinformatic analyses were executed by M.B., L.F., Y.P., N.D. and Y.H. G.L., M.B., H.-J.M. and O.M. wrote the report, with input from L.F., B.H. and M.A.M.G.

## Additional information

**Accession codes:** All BAM files for the 70 re-sequenced samples have been deposited in the European Nucleotide Archive (ENA) under the accession code ERP001813.

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://www.nature.com/reprintsandpermissions/>

**How to cite this article:** Bosse, M. et al. Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nat. Commun.* 5:4392 doi: 10.1038/ncomms5392 (2014).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>