nature genetics

Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing

Huashui Ai[1,4], Xiaodong Fang[2,4], Bin Yang[1,4], Zhiyong Huang[2], Hao Chen[1], Likai Mao[2], Feng Zhang[1], Lu Zhang[2], Leilei Cui[1], Weiming He[2], Jie Yang[1], Xiaoming Yao[2], Lisheng Zhou[1], Lijuan Han[2], Jing Li[1], Silong Sun[2], Xianhua Xie[1], Boxian Lai[2], Ying Su[1], Yao Lu[2], Hui Yang[1], Tao Huang[1], Wenjiang Deng[1], Rasmus Nielsen[2,3], Jun Ren[1,5] & Lusheng Huang[1,5]

Domestic pigs have evolved genetic adaptations to their local environmental conditions, such as cold and hot climates. We sequenced the genomes of 69 pigs from 15 geographically divergent locations in China and detected 41 million variants, of which 21 million were absent from the dbSNP database. In a genome-wide scan, we identified a set of loci that likely have a role in regional adaptations to high- and low-latitude environments within China. Intriguingly, we found an exceptionally large (14-Mb) region with a low recombination rate on the X chromosome that appears to have two distinct haplotypes in the high- and low-latitude populations, possibly underlying their adaptation to cold and hot environments, respectively. Surprisingly, the adaptive sweep in the high-latitude regions has acted on DNA that might have been introgressed from an extinct *Sus* species. Our findings provide new insights into the evolutionary history of pigs and the role of introgression in adaptation.

Animal domestication is one of the most important events in human history, allowing a transition from hunting and gathering to more settled lifestyles. Pigs were domesticated largely in the Near East and China, approximately 10,000 years ago[1,2]. Since then, pigs have been subject to the combined effects of natural selection and human-driven artificial selection, resulting in marked phenotypic diversity in appearance, fertility, growth, palatability and local fitness[3]. China is now the leading country in terms of genetic resources for domestic pigs, having more than one-third (~100 breeds) of the total number of global breeds. Chinese indigenous breeds have evolved genetic adaptations to various environmental conditions in the vast geographical region of China[3]. For example, local pigs from southern and northern China have distinct thermoregulatory mechanisms for hot and cold temperatures in low- and high-latitude areas, respectively, providing an exceptional opportunity to elucidate the genetic basis of adaptive evolution, which remains largely unexplored[3]. However, identifying the causal genetic variants underlying naturally adapted traits is challenging. Only a handful of such variants have been unambiguously identified in humans and other organisms[4–8]. Recently, whole-genome sequencing of representative individuals from diverse populations has become feasible. Population genomics studies that conducted genome scans for regions shaped by selection have been used to efficiently characterize candidate genes that contribute to phenotypic diversity in both model organisms and domestic animals[9–12]. In pigs, whole-genome sequencing and selective sweep analysis have been used to identify the genomic signatures of selection contributing to the domestication of European pigs and the local adaptation of the Chinese Tibetan wild boars[13–15]. Here we sequenced the genomes of 69 indigenous Chinese pigs from high- and low-latitude environments at high genomic coverage (>25-fold), enabling us to compile a nearly complete catalog of genetic variants, identify a genome-wide set of candidate loci for local adaptation in Chinese pigs and provide new insights into the adaptive evolutionary history of the pig.

RESULTS

We selected 69 individuals to represent 11 geographically diverse breeds and 3 populations of wild boar (**Fig. 1a** and **Supplementary Fig. 1**) from cold and hot environments in China (**Supplementary Table 1** and **Supplementary Note**). We performed whole-genome sequencing for the 69 pigs. Two libraries with insert sizes of 500 bp were constructed for each individual and sequenced using the HiSeq2000 platform (Illumina). We used the Wuzhishan pig sequence[16] as the reference genome for its better alignment with Chinese pig sequences in comparison to the international Duroc assembly[13] (**Supplementary Tables 2** and **3**, and **Supplementary Note**). The sequence data for each individual reached more than 25-fold depth and 95% genome coverage (**Table 1**), allowing us to call variants with high confidence.

Characterization of variants

After applying stringent quality control criteria (Online Methods), we identified a total of 40,820,483 SNPs in the 69 genomes (**Table 1**),

[1]Key Laboratory for Animal Biotechnology of Jiangxi Province and the Ministry of Agriculture of China, Jiangxi Agricultural University, Nanchang, China. [2]BGI-Tech, BGI-Shenzhen, Shenzhen, China. [3]Department of Integrative Biology, University of California, Berkeley, Berkeley, California, USA. [4]These authors contributed equally to this work. [5]These authors jointly directed this work. Correspondence should be addressed to J.R. (renjunjxau@hotmail.com) or L.Huang (lushenghuang@hotmail.com).

Received 20 August 2014; accepted 29 December 2014; published online 26 January 2015; doi:10.1038/ng.3199

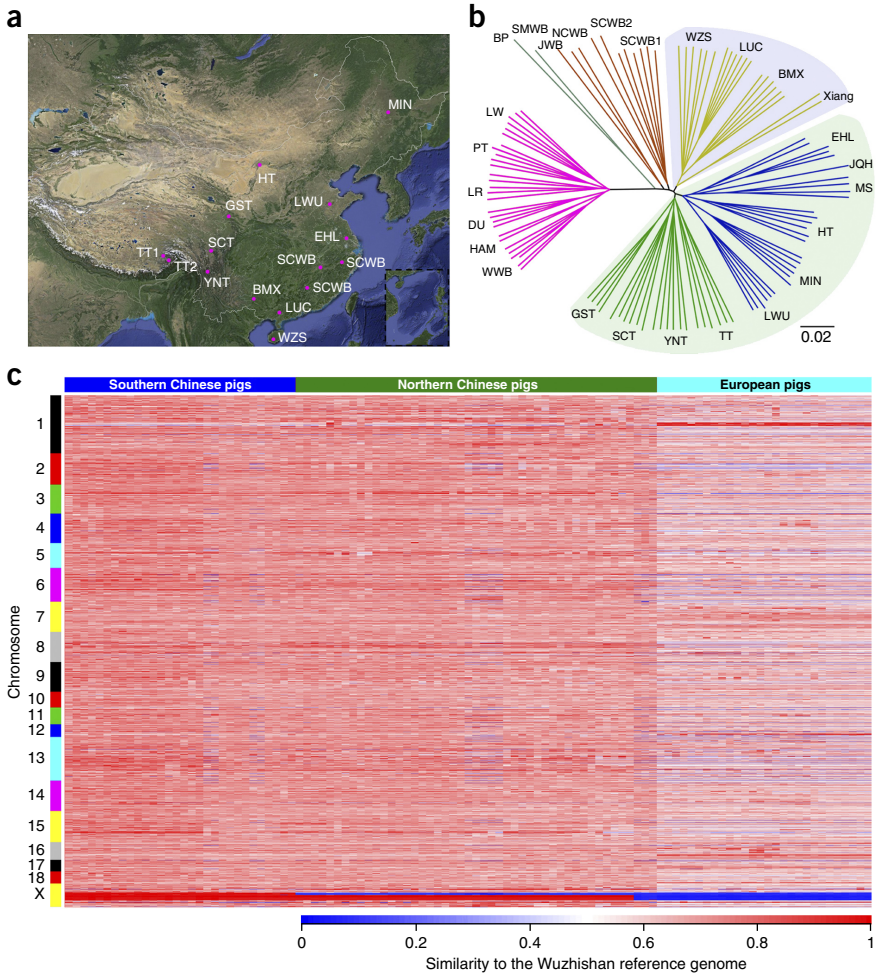

Figure 1 Genomic variation in Chinese pigs. (**a**) Geographical distribution of the indigenous Chinese pigs used in our study. (**b**) Neighbor-joining tree of Chinese and European pigs based on our data and publicly available whole-genome sequences of pigs. Different colors represent clusters of subpopulations geographically close to each other. Light-green shading indicates northern Chinese domestic pigs, and light-blue shading indicates southern Chinese domestic pigs. The scale bar represents the identity-by-state (IBS) score between individuals. (**c**) Genomic similarity of Chinese and European pigs to the Wuzhishan reference genome. Chromosomes are indicated by different colors along the left *y* axis. Identical score (IS) values are shown for SNPs within each 50-kb window across the genome. BP, bearded pig (*S. barbatus*); BMX, Bamaxiang; DU, Duroc; EHL, Erhualian; GST, Tibetan (Gansu); HAM, Hampshire; HT, Hetao; JQH, Jiangquhai; JWB, Japanese wild boar; LR, Landrace; LW, Large White; LWU, Laiwu; LUC, Luchuan; MIN, Min; MS, Meishan; NCWB, northern Chinese wild boar; PT, Pietrain; SCT, Tibetan (Sichuan); SCWB1, southern Chinese wild boars sequenced in this study; SCWB2, southern Chinese wild boars for which genome sequences are publicly available; SMWB, Sumatran wild boar; TT, Tibetan (Tibet); YNT, Tibetan (Yunnan); WZS, Wuzhishan. Whole-genome data for BP, DU, HAM, JQH, JWB, LR, LW, MS, NCWB, PT, SCWB2 and SMWB were from Groenen *et al.*[13].

of which 26.6 million were intergenic, 0.5 million were intronic and 188,664 were exonic (**Supplementary Table 4**). To assess sequence accuracy, we compared the SNP calls for the whole-genome sequencing to the SNPs on the porcine 60K BeadChip genotyping array (Illumina) for which data were available for the 69 pigs. Of the 53,982 polymorphic loci in the 60K data set, more than 98% (53,280) were consistent with the SNPs identified from sequencing data, demonstrating the high quality and reliability of our SNP calls (**Supplementary Fig. 2**).

We next compared the SNPs that we identified with those from Build 138 of the pig dbSNP database. Over 70% of the variants (19,508,676 SNPs) in the dbSNP database were found in our SNP data set, whereas more than half (>52%; 21,311,807 SNPs) of the variants that we identified were absent from the dbSNP database (**Supplementary Fig. 3**). These novel SNPs substantially expand the catalog of porcine genetic variants.

Consistent with previous findings[13,17], we observed the highest number of SNPs per individual in the Chinese wild boars, with 14.5 million SNPs per individual, corresponding to ~9% more SNPs than the average of 13.3 million (**Table 1**). At the genome level, the termini of chromosomes exhibited higher nucleotide variability (**Supplementary Figs. 4** and **5**). This finding is in agreement with a previous observation[18], and this increase in variability could be caused by elevated recombination rates at the ends of chromosomes[18,19]. A total of 25.7 million (62.9%) SNPs had a minor allele frequency (MAF) greater than 0.05 (**Supplementary Fig. 6**). Wild boars had the largest number (767,770) of population-specific SNPs (**Table 1**).

We detected an average of 794 putative loss-of-function mutations in each Chinese pig breed (**Supplementary Table 5**). We found only three breed-specific loss-of-function variants: a nonsense mutation in the *PKD1L3* gene in Laiwu pigs, a splice-site mutation in the *IFLTD1* gene in Luchuan pigs and a splice-site mutation in the *SKIL* gene in Min pigs (**Supplementary Table 5**). These specific loss-of-function variants might have had a role in the formation of the characteristic phenotypes of these breeds.

We also identified 5,663,829 indels and 44,170 structural variations (**Table 1**) that were similarly distributed in the 69 genomes (**Supplementary Fig. 7**). Transposable elements are a major source of structural variation between individuals in humans[20]. The Wuzhishan reference genome contains more than 2 million copies of the tRNA-derived short interspersed element (SINE/tRNA), a common transposable element in mammals[16]. To examine whether SINE/tRNA elements are also a source of structural variation in the pig genome, we investigated the size distributions of the detected structural variations and SINE/tRNA elements. We found that structural variations and SINE/tRNA elements had similar distribution patterns—more than 50% of structural variations overlapped with SINE/tRNA elements (**Supplementary Fig. 8** and **Supplementary Table 6**). This observation supports the idea that SINE/tRNA elements are also an important source of genetic diversity in pigs.

To determine the genomic similarity of Chinese and European pigs, we downloaded the publicly available whole-genome sequence data for 26 European and 16 Asian pigs, including 13 Chinese pigs (**Supplementary Table 7**). We combined these data with our data to create a 111-individual SNP data set comprising 40,334,686 SNPs (Online Methods). Population genetics analysis based on the

Table 1 Summary statistics for the whole-genome sequences

							Tibetan							
	Bamaxiang	Erhualian	Hetao	Laiwu	Luchuan	Min	Gansu	Sichuan	Tibet	Yunnan	Wuzhishan	Wild boar[a]	Overall	Average
Number	6	5	6	6	6	6	4	6	6	6	6	6	69	–
Average sequencing depth	26.9	34.0	24.3	26.2	23.8	25.7	25.6	27.0	24.4	26.9	26.7	26.4	317.9	26.5
Average genome coverage (%)	95.1	88.6	95.0	95.3	95.1	95.1	95.1	95.4	95.5	95.6	95.2	95.2	–	94.7
SNPs	17,995,352	15,146,774	17,839,365	16,544,235	14,872,507	16,904,500	15,066,167	18,920,622	19,739,516	19,413,138	19,622,595	21,317,186	40,820,483	17,781,830
SNPs per individual	13,391,933	11,756,968	14,064,290	13,666,968	12,507,483	14,103,695	11,008,269	13,654,403	13,876,078	14,080,417	13,529,363	14,554,303	–	13,349,514
Coding SNPs	107,332	98,475	97,624	97,415	83,135	98,415	77,689	109,494	111,430	114,378	111,039	117,729	–	102,013
Fixed SNPs	882,221	1,731,526	1,276,699	1,804,768	1,211,676	1,769,696	1,113,806	1,195,753	897,321	947,718	546,802	741,818	–	1,176,650
Unique SNPs	215,191	124,335	219,556	194,831	127,937	257,350	122,879	242,046	302,275	259,971	206,518	767,770	–	253,388
Indels	2,557,636	2,204,152	2,544,473	2,392,951	2,176,771	2,445,718	2,176,401	2,709,053	2,770,629	2,760,314	2,785,251	3,070,303	5,663,829	2,549,471
Coding indels	2,146	2,143	1,941	1,972	1,874	2,000	1,697	2,130	2,151	2,216	2,261	2,387	–	2,077
Insertions per individual	679,892	573,379	685,557	698,023	633,557	716,151	674,388	699,950	664,971	702,518	683,568	714,358	–	677,193
Deletions per individual	706,507	592,103	722,847	726,967	653,950	748,081	709,189	728,679	690,689	733,925	708,991	747,784	–	705,809
Structural variations	35,711	30,558	36,343	38,101	34,717	39,569	29,917	37,554	40,916	37,144	36,757	40,216	44,170	36,459
Structural variations per individual	21,262	15,477	21,696	23,833	20,753	23,202	20,103	21,868	22,501	23,803	20,622	21,764	–	21,497
Nucleotide diversity (%)	0.43	0.38	0.47	0.38	0.36	0.39	0.40	0.45	0.49	0.47	0.48	0.51	–	0.43

[a]Wild boars sequenced in this study are from southern China.

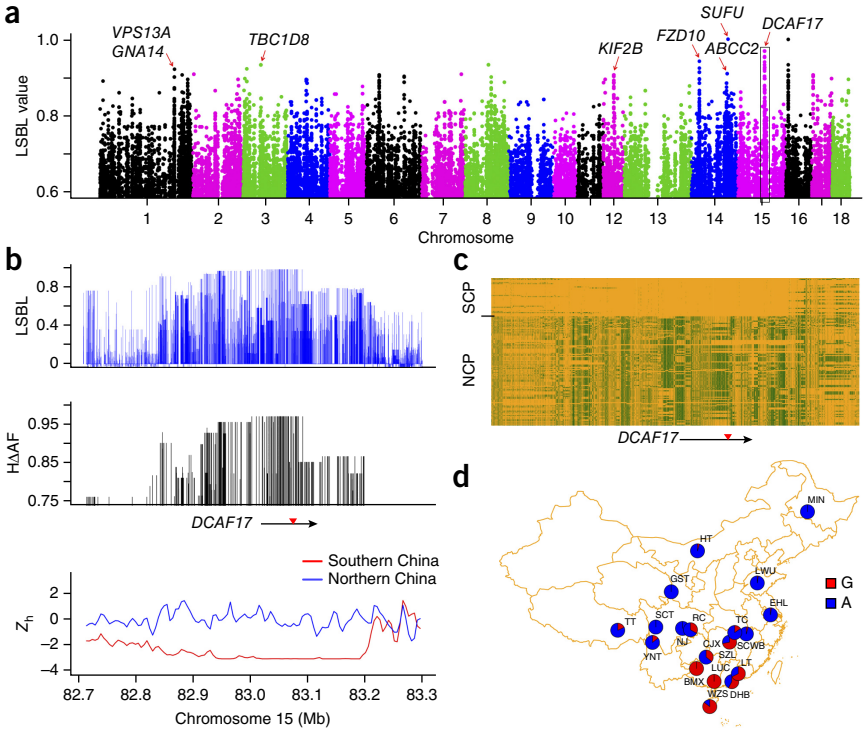

Figure 2 Identification of candidate genes for local adaptation on the autosomes. (**a**) Divergent genomic loci between southern and northern Chinese pigs. LSBL values for 50-kb windows are plotted along pig autosomes 1–18. Chromosomes are represented by color. Each dot represents a 50-kb window. The peak on chromosome 15 (boxed) consists of multiple dots and includes the *DCAF17* gene. The candidate genes corresponding to the top ten outlier regions are labeled on the plot. (**b**) The three statistics of LSBL, SNPs with ΔAF > 0.75 (HΔAF) and Z_h support a strong sweep signal around the chromosome 15 region. The red triangle indicates the position of the top outlier SNP representing the *DCAF17* gene. (**c**) The degree of haplotype sharing in pairwise comparisons between breeds around the chromosome 15 region. Haplotype sharing is much more extensive in the pig breeds and wild boars from southern China (SCP) than in those from northern China (NCP). The major allele in SCP (NCP) is indicated by orange (green). (**d**) Allele distribution of the top outlier SNP representing the *DCAF17* gene in 462 Chinese domestic pigs from 18 diverse pig breeds and 25 wild boars. SNP genotypes were determined by Sanger sequencing. Red (blue) color represents the major allele in pigs from southern (northern) China. CJX, Congjiangxiang; DHB, Dahuabai; HUAI, Huai; LT, Lantang; NJ, Neijiang; RC, Rongchang; SZL, Shaziling; TC, Tongcheng; WB, wild boar. The other breed abbreviations are as in **Figure 1**.

111-genome SNP data set showed a clear evolutionary split and reciprocal introgression between Chinese and European pigs (**Supplementary Figs. 9–14**, **Supplementary Tables 8–10** and **Supplementary Note**). A neighbor-joining tree (**Fig. 1b**) showed this divergence: the European pigs defined their own separate clade, supporting previous reports of the independent domestication origins of Chinese and European pigs[1,2]. Identity score (IS) analysis clearly illustrated that European pigs had less genomic similarity to the Wuzhishan reference genome than Chinese pigs (**Fig. 1c**).

Selective sweeps on autosomes

To identify genomic loci that favor local adaptation to hot or cold environments in Chinese high-latitude (hereafter referred to as northern Chinese) and low-latitude (southern Chinese) pigs, we first performed locus-specific branch-length (LSBL) analyses (Online Methods) using all of the called SNPs in the 69 genomes. Using stringent criteria (Online Methods), we identified a total of 774 putative sweep regions with an average size of 51 kb on the autosomes (**Supplementary Table 11**) and a large sweep region of ~14 Mb on the X chromosome. A majority of the outlier SNPs were intergenic or intronic, suggesting that regulatory variants might have had prominent roles in genetic adaptations to local environments for Chinese pigs.

On the autosomes, we identified 219 genes corresponding to selective sweeps (**Supplementary Table 12**). Gene Ontology (GO) analysis identified a significant over-representation of genes involved in biological processes that contribute to the maintenance of thermostatic status during heat or cold stress. These processes were related to hair development, forebrain neuron differentiation, kidney development, energy metabolism and blood circulation, among others (**Supplementary Table 13**). For example, we found enriched genes involved in hair cell differentiation (*ATOH1*, *JAG1* and *RAC1*; $P = 0.02$) and hair follicle maturation (*BARX2* and *TBC1D8*; $P = 0.02$). This finding is consistent with the fact that southern Chinese pigs have sparse, short hair beneficial for heat loss and northern Chinese pigs generally have long, dense hair serving as an insulating layer. We observed the enrichment of genes participating in forebrain neuron differentiation (*DLX1*, *DLX2*, *RAC1*, *ROBO1* and *SALL1*; $P = 0.003$). This finding is in agreement with the knowledge that the central nervous system has an important role in temperature acclimation by invoking a first-line heat loss or production response[21]. We noted over-represented biological processes related to kidney development (*BMP4*, *BMP7*, *MYC*, *SALL1*, *SPRY1* and *KLHL3*; $P = 0.01$). In humans, cold stress can increase blood pressure and urine production and reduce renal water reabsorption[22]. Kidney weight tends to increase in the cold and decrease in the heat in several species[23]. We also found enriched functional categories related to blood circulation, including artery development (*BMP4*, *CITED2* and *JAG1*; $P = 0.02$) and embryonic heart tube development (*CITED2*, *INVS*, *RYR2*, *SUFU* and *TBC1D8*; $P = 0.02$). It has been reported that heat acclimation is achieved by increasing skin blood flow, which increase heat loss[24]. In rats adapting to a warm temperature, blood flow increases in muscular organs and the adrenals, whereas, in rats acclimating to cold, blood flow increases in adipose tissues, the kidney, intestines and the liver[25]. Therefore, our results illustrate the important role of biological pathways influencing blood flow in the temperature adaptation of Chinese pigs.

Most of the genes located in the top ten most significant sweep regions (**Fig. 2a**) are functionally plausible for temperature adaptation, according to their annotations in the NCBI Gene database. These genes included *SUFU* affecting neural tube closure and skin development; *TBC1D8*, *VPS13A*, *GNA14* and *KIF2B* associated with blood coagulation and circulation; *DCAF17* (also known as *C2orf37*) affecting hair development; *FZD10* regulating vasculature development; and *ABCC2* associated with cellular chloride ion homeostasis and response to heat. These genes are appealing candidates for further investigation.

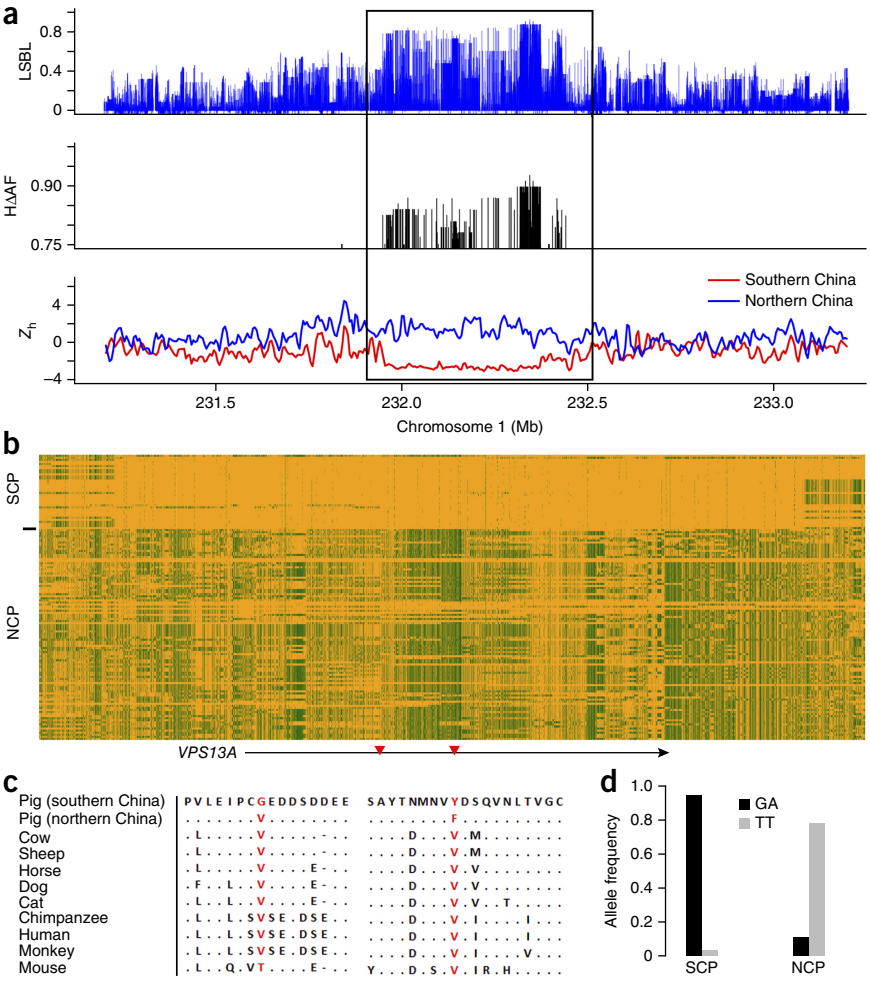

Figure 3 The selective sweep region around the *VPS13A* gene. (**a**) Three sweep statistics plotted over a ~2-Mb region on chromosome 1. From top to bottom, the vertical axis shows the values of LSBL, the proportions of SNPs with ΔAF > 0.75 (HΔAF) and Z_h. The statistics were calculated separately for the northern (high-latitude including Tibet) and southern (low-latitude) Chinese pigs. These statistics clearly indicate a strong sweep signal around the *VPS13A* gene that exhibits strong LSBL scores between southern and northern Chinese pigs, divergent allele frequency and reduced nucleotide diversity. The sweep region is highlighted with a rectangle. (**b**) The degree of haplotype sharing in pairwise comparisons between populations. The major allele in SCP (NCP) is indicated by orange (green). Haplotype sharing is much more extensive in the pig breeds and wild boars from southern China than in those from northern China. (**c**) The multi-species alignment of two nonsynonymous substitutions encoded in the *VPS13A* gene (red triangles in **b**). The amino acid sequences in European pigs around the two substitutions shown here are the same as those in northern Chinese pigs. Dots indicate identity with the master sequence, and dashes indicate missing data. (**d**) Comparison of the haplotype frequencies at the two protein-altering sites in 18 southern Chinese pigs and 34 northern Chinese pigs.

We found the most prominent LSBL signature on chromosome 15 (**Fig. 2a**), which appeared to be a strong selective sweep spanning a 300-kb region (82.9–83.2 Mb). The signature was further supported by statistics of the absolute allele frequency difference (ΔAF) between southern and northern Chinese pigs (**Fig. 2b**), standardized heterozygosity (Z_h) (**Fig. 2b**) and the degree of haplotype sharing (**Fig. 2c**). The 300-kb region was nearly devoid of genetic variability, enriched for SNPs with high ΔAF (>0.75) and showed long-range haplotype sharing in southern Chinese pigs, indicative of a strong selective sweep. A functionally plausible gene in this region was *DCAF17*, which has been implicated in human Woodhouse-Sakati syndrome characterized by hair loss and other symptoms[26]. We genotyped the top LSBL SNP representing the *DCAF17* gene in 462 Chinese domestic pigs from 18 diverse breeds and 25 wild boars. The top SNP showed clear allelic imbalance between northern and southern Chinese pigs (**Fig. 2d**). These finding support the hypothesis that *DCAF17* likely has a role in local adaptation by regulating hair development in southern Chinese pigs.

Next, we searched for missense mutations representing putative functional variants in the 219 candidate genes. Altogether, we identified 15 protein-altering SNP outliers in terms of LSBL values in 10 genes (**Supplementary Table 14**). Notably, two nonsynonymous substitutions were present at conserved sites within the *VPS13A* gene. These SNPs displayed marked allele frequency differences between northern and southern Chinese pigs (**Fig. 3**). The *VPS13A* region also showed strong evidence of reduced heterozygosity (**Fig. 3a**), enrichment for SNPs with a large difference in allele frequency relative to northern Chinese pigs (**Fig. 3a**) and long-range linkage disequilibrium (LD) in southern Chinese pigs (**Fig. 3b**). The *VPS13A* protein-altering mutations that colocalized with the sweep region probably have functional relevance and have undergone directional selection for heat adaptation in southern Chinese pigs. *VPS13A* encodes chorein, recently characterized as a key regulator of the secretion and aggregation of blood platelets[27]. In humans, platelet counts affect whole-blood viscosity. Heat stress can increase platelet counts and blood viscosity, which in turn increase the risk of cerebral and coronary thrombosis[28]. This led us to hypothesize that the *VPS13A* missense mutations might contribute to reducing the risk of thrombosis by modulating platelet counts and blood viscosity in southern Chinese pigs in hot environments.

An exceptionally large sweep region on the X chromosome

We were particularly interested in the X-linked sweep region (**Fig. 4**) that has also been observed in European pigs[14] because the selection signal was extremely strong within an exceptionally large region (14 Mb in length): a total of 84,373 LSBL outlier SNPs were enriched in the 14-Mb region (44.0–57.8 Mb) on the X chromosome, accounting for 75.2% of all outliers. Of the 74,515 SNPs with extreme differences in allele frequency between the northern and southern Chinese pigs (with frequencies >90% in one group and <10% in the other), 94.5% were present in this region, which showed unusually long-range (14-Mb) complete LD. Despite the phylogenetic split (**Fig. 1b**) and divergent appearance of the Tibetan pigs and other northern Chinese breeds, the strong selection signal in this region was shared by all northern Chinese breeds, including Tibetan pigs. All of the northern Chinese individuals had a core haplotype that was distinct from those of southern Chinese pigs (**Fig. 4a**). Furthermore, this region exhibited unusually low heterozygosity in both northern and southern Chinese pigs

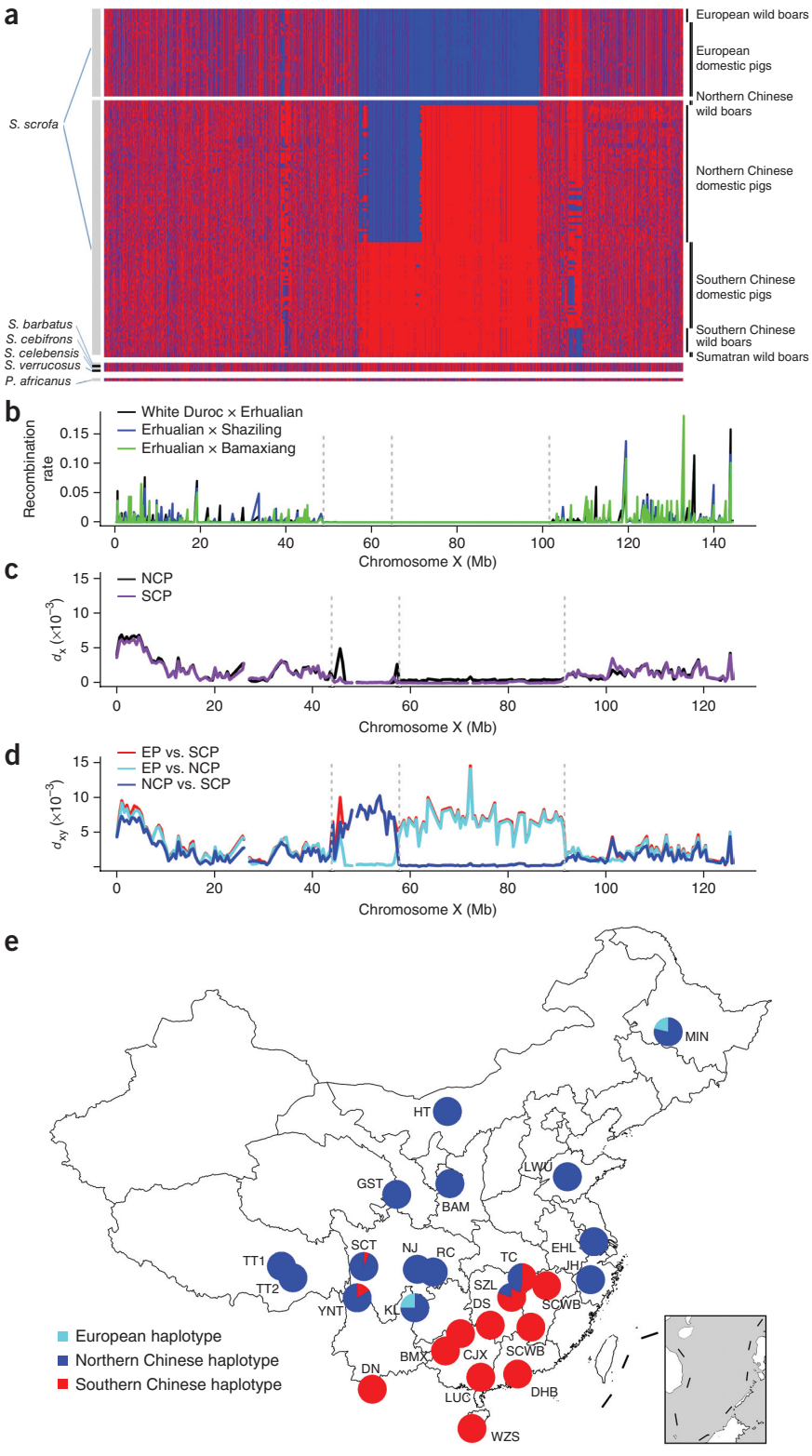

Figure 4 Characterization of the X-linked selective sweep region. (**a**) The pattern of haplotype sharing in diverse populations. The haplotypes were reconstructed for each individual using all of the variants on the X chromosome. Alleles that are identical to or different from the ones in the Wuzhishan reference genome are indicated by red and blue, respectively. (**b**) The distribution of recombination rates on the X chromosome. The vertical gray dashed lines (also in **c** and **d**) denote the boundaries of the 14-Mb sweep and its flanking 34-Mb region in the recombination coldspot. (**c**) The plot for the d_x statistic (the number of pairwise differences per site) in a window size of 500 kb on chromosome X within northern Chinese domestic pigs (NCP) and within southern Chinese wild boars and domestic pigs (SCP). (**d**) The plot for the d_{xy} statistic (the number of pairwise differences per site) in a window size of 500 kb on chromosome X between three contrasting groups, including European wild boars and domestic pigs (EP) versus SCP, EP versus NCP, and NCP versus SCP. (**e**) The geographical distribution of three haplotypes corresponding to the 48-Mb region with a low recombination rate in Chinese pigs. The haplotypes were phased using seven tagging SNPs on the porcine 60K Chip (Illumina) within the 48-Mb region. BAM, Bamei; DN, Diannan; DS, Dongshan; JH, Jinhua; KL, Kele. The other breed abbreviations are as in **Figures 1** and **2**.

(**Supplementary Figs. 15** and **16**). Notably, unlike classic sweeps that usually exhibit signatures of selection in a certain group within relatively small regions, the X-linked sweep was exceptionally large with signatures in both northern and southern Chinese pigs.

Strong signals of selection on the X chromosome are expected because recessive mutations are exposed to selection in hemizygous males. However, the length of the sweep region is unusual (**Fig. 4a**). The extreme length of the region, spanning 14 Mb, is explained by the fact that it is located in a 48-Mb segment (44.0–91.5 Mb) of chromosome X that has an extremely low rate of recombination[29] (**Fig. 4b**). Because the size of a sweep region is inversely proportional to the recombination rate[30], the large length of this sweep region is not unexpected.

Low recombination rate in the X-linked sweep region

We carried out further investigation to clarify the mechanism underlying the extremely low recombination rate in the 48-Mb region on chromosome X. Large-scale structural variants such as inversions can suppress recombination locally. If all individuals from one group, such as northern Chinese pigs, were fixed for a structural variant, we would expect to observe recombination events within this group. To test this hypothesis, we examined the recombination rate along chromosome X in three F_2 intercross populations: White Duroc (European breed) × Erhualian (northern Chinese breed), Erhualian × Shaziling (northern Chinese breeds) and Erhualian × Bamaxiang (southern Chinese breeds). However, our pedigree analysis showed that recombination was strongly suppressed both within *Sus scrofa* populations from northern China (Erhualian × Shaziling) and hybrid populations of northern and southern Chinese breeds (Erhualian × Bamaxiang) (**Fig. 4b**). We hence argue that structural variation is not likely to be the cause of the observed low recombination rate.

We noted that this region encompassed the centromere (47.3–49.2 Mb), a structure that usually shows a low recombination rate in a wide range of organisms, including humans (**Supplementary Fig. 17**)[31]. We further observed significantly lower GC content ($P < 1.3 \times 10^{-72}$) and greater repeat sequence extent ($P < 4.0 \times 10^{-32}$) in this region than for the remainder of the sex chromosome (**Supplementary Fig. 18** and **Supplementary Table 15**). By aligning a 2-Mb sequence at each boundary site of the 48-Mb region to the entire X chromosome, we identified the enrichment of a 6-kb poly(T) core sequence in the 48-Mb region ($P < 1.4 \times 10^{-28}$; **Supplementary Fig. 19**). The poly(T) sequence was negatively correlated (Spearman's correlation, $P < 1 \times 10^{-16}$) with the recombination rate at a fine scale on autosomes (**Supplementary Table 16**), consistent with the negative correlation between poly(T)-rich sequences and recombination in human genomes[32].

There were three major groups of haplotypes in the 48-Mb region: one in European pigs (domestic breeds and wild boars) and northern Chinese wild boars, a highly differentiated one in southern Chinese pigs (domestic breeds and wild boars) and a third haplogroup in northern Chinese domestic pigs, which appeared to be a recombinant between the other two haplogroups (**Fig. 4a**). We further examined tagging SNPs within the 48-Mb region on the porcine 60K chip (Illumina) in a large sample of 422 Chinese pigs and confirmed the recombinant haplogroup in northern Chinese domestic pigs (**Fig. 4e** and **Supplementary Table 17**). The recombinant haplogroup may suggest that northern Chinese domestic pigs arose through hybridization of northern and southern *S. scrofa* varieties, either before or after domestication. The recombinant haplogroup comprised two segments of 14 Mb and 34 Mb in length (**Fig. 4a**). Notably, the GC content appeared to be higher and the matches to poly(T) sequence were less prevalent at the border region between the two segments (**Supplementary Figs. 18** and **19**). This finding is in agreement with the strong correlation between GC-rich sequences and a high recombination rate in the pig genome[18] and further supports the enrichment of poly(T) sequence as a cause for reduction in the recombination rate. Altogether, we propose that poly(T)-enriched sequences, including the 6-kb core repeat, are potential factors suppressing recombination events in the 48-Mb region.

The X-linked sweep region is caused by natural selection

Surprisingly, there was a strong reduction in heterozygosity in the 48-Mb region in both northern and southern Chinese pigs (**Fig. 4c**, **Supplementary Figs. 15** and **16**, and **Supplementary Table 18**), even though the northern and southern Chinese haplotypes were highly divergent from each other. We note that this pattern cannot be explained by a general reduction in variability, for example, due to a reduced mutation rate, as the region did not show signs of a reduced mutation rate in comparisons of African warthogs (*Phacochoerus africanus*) and *Sus* species. African warthogs had similar nucleotide distances (d_{xy}) to all *Sus* populations in both the sweep region and the remainder of the X chromosome (**Fig. 5a**). The comparisons of *S. scrofa* with other *Sus* species indicated that it had a twofold reduction in variability, as reflected by d_{xy} values. In contrast, northern and southern Chinese pigs had a 10- to 15-fold reduction in nucleotide variability in the sweep region relative to the remainder of the X chromosome (**Fig. 5a**). Moreover, the reduction in variability within both northern and southern Chinese pigs was so extreme that it is unlikely to have been caused by a lack of recombination and resulting high coalescence variance but insteadis likely to be explained by natural selection. This hypothesis is supported by coalescence simulations that showed that such a reduction in variability was unlikely to be observed through random chance, even in the complete absence of recombination (**Supplementary Tables 19** and **20**, and **Supplementary Note**). Altogether, these observations provide strong evidence that more than one selective sweep has affected the region, reducing variability in both northern and southern Chinese pigs, although different haplotypes are affected.

We approximated the sweep ages for the 14-Mb haplotype by calculating the average pairwise time to the most recent common ancestor (t_{MRCA}) of the haplotype in northern and southern Chinese pigs separately. The estimate of ~0.13 million years (**Supplementary Table 21**) corresponds to the beginning of the last glacial period (0.13–0.01 million years ago). The global glaciations during this time period might have imposed severe challenges, especially in organisms living at high latitudes[33], and consequently caused bottlenecks in the population size as reported previously[13]. The selection footprints we detected likely date back to this time and were subsequently incorporated into present-day domestic pigs.

Possible ancient interspecies introgression

We were puzzled at the extreme divergence between the northern and southern haplotypes in the 14-Mb region (**Figs. 4d** and **5a**). After applying a molecular clock model (Online Methods), we estimated that they diverged ~8.5 million years ago (**Supplementary Table 22**), an estimate that far exceeds the average coalescence time in the pig genome (**Supplementary Table 22**), meaning that their divergence long preceded the known evolutionary history from ~5 million years ago of *S. scrofa*[34]. One explanation for this odd phenomenon is that the two distinct haplotypes originated, respectively, from two subpopulations of an ancestral Suidae family from which *S. scrofa* and other *Sus* species evolved. The extremely low recombination rate could have maintained divergence between the two haplotypes, making the average coalescence time in this region much larger than those for the other regions in the *S. scrofa* genome. If so, it is really unprecedented that the polymorphisms in this region have been maintained for 8.5 million years. Further work is needed to explicitly assess support for this hypothesisof an ancient substructure or introgression.

Another reasonable hypothesis is that the sweep haplotype was introgressed from other species closely related to *S. scrofa*. To explore the introgression hypothesis, we further analyzed whole-genome sequence data for one *Phacochoerus*species (African warthog; *P. africanus*) and four *Sus* species, including bearded pig (*Susbarbatus*), Celebes warty pig (*Sus celebensis*), Java warty pig (*Sus verrucosus*) and Visayan warty pig (*Sus cebifrons*), from Frantz *et al.*[34]. The northern haplotype was distinct from all other *Sus* haplotypes (**Fig. 4a**). We reconstructed the phylogenetic trees for the 14-Mb region, an autosomal region of equivalent size (14 Mb) on chromosome 2 and the 34-Mb region on chromosome X (**Fig. 5b**). The phylogenetic pattern in the tree for the autosomal region was consistent with the known evolutionary history of *Sus* species, with the warthog appearing as a clear outgroup to all members of the *Sus* genus. However, in the tree for the 14-Mb region, European pigs and northern Chinese pigs formed a separate clade. This clade had strong bootstrap support and was separated with a very long branch from other members of the *Sus* genus. African warthog again appeared as a divergent outgroup. The tree for the 34-Mb region was nearly identical to that for the 14-Mb region, except that northern Chinese domestic pigs clustered with southern Chinese domestic pigs and wild boars. In this tree, European pigs and northern Chinese wild boars again formed a separate clade, which was very divergent from other members of the *Sus* genus. These results support the introgression hypothesis and indicate that the northern haplotype was likely introgressed from a (possibly) extinct

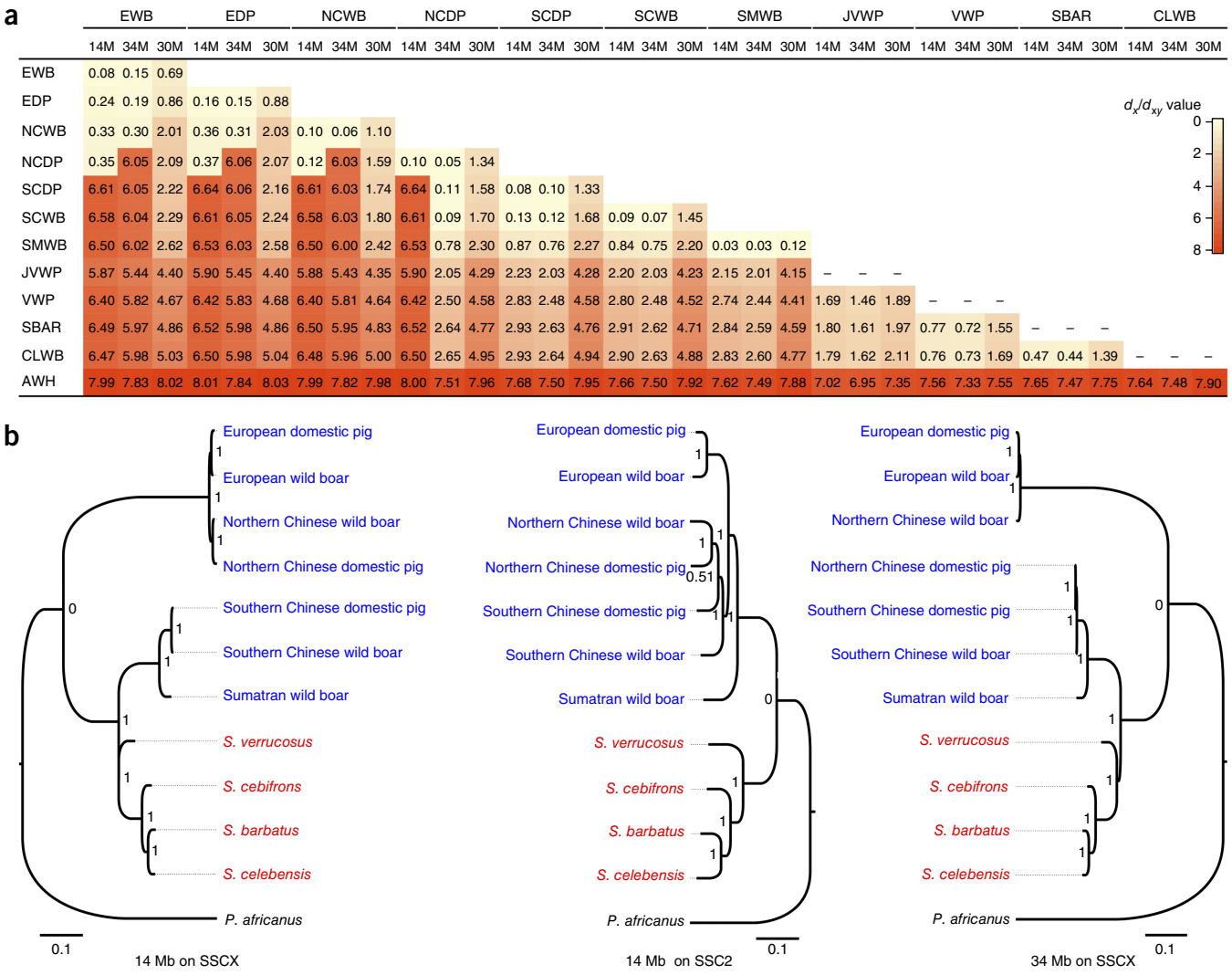

Figure 5 The X-linked selective sweep in northern Chinese pigs was possibly introgressed from an extinct *Sus* species. (**a**) Nucleotide distance within (d_x) and between (d_{xy}) pig breeds (1,000×). d_x and d_{xy} values are highlighted with different density colors, where darker colors correspond to larger values. 14M, the 14-Mb sweep region on chromosome X; 34M, the 34-Mb region with a low recombination rate flanking the 14-Mb region on chromosome X; 30M, the 30-Mb region (14–44 Mb) flanking the 14-Mb region on chromosome X. EWB, European wild boars; EDP, European domestic pigs; NCWB, northern Chinese wild boars; NCDP, northern Chinese domestic pigs; SCDP, southern Chinese domestic pigs; SCWB, southern Chinese wild boars; SMWB, Sumatran wild boars; JVWP, *S. cebifrons*, VWP, *S. verrucosus*; SBAR, *S. barbatus*; CLWB, *S. celebensis*; AWH, African warthog. (**b**) The maximum-likelihood trees for different genomic regions, including the 14-Mb region and its flanking 34-Mb region on chromosome X and an autosomal region of 14 Mb (44.0–57.8 Mb) on chromosome 2. Scale bars represent the number of nucleotide substitutions per SNP site.

Sus suid, which constitutes as an outgroup to *S. scrofa* and the four *Sus* species confined to Island Southeast Asia. Surprisingly, the northern haplotype, possibly derived from a divergent species in another genus, must have spread through northern Chinese and European wild boars, as this haplotype was present in both populations (**Fig. 4a**). This spread might have occurred before the Asian-European split of wild boars ~1.2 million years ago or might have been caused by gene flow between wild boars of northern China and Europe during the Pleistocene era after their divergence, a hypothesis supported by an excess of derived similarity between European and northern Chinese wild boars on the autosomes as reported previously[13].

DISCUSSION

We used high-coverage whole-genome sequencing to generate a comprehensive catalog of genetic variants in a sample containing a broad panel of 69 Chinese pigs. This is one of the first population genomics analyses to use high-coverage whole-genome sequencing in pigs. The data represent a valuable resource for evolutionary analyses, in particular, for identifying functionally important variants contributing to the phenotypic diversity of Chinese pigs. We identified a genome-wide set of candidate targets of natural selection for local adaptation to environments at varying latitudes in the diverse Chinese pig breeds. Notably, we discovered an exceptionally large selective sweep region on the X chromosome that appears to have undergone extremely strong selection in both northern and southernChinese populations. Surprisingly, the adaptive haplotype in the northern Chinese populations was likely introduced from another divergent *Sus* species, providing the first evidence, to our knowledge, that interspecies introgression has driven adaptation in a mammal. Our ability to detect this potentially quite old introgression event is facilitated by the fact that the introgression fragment falls in a region with unusually reduced recombination rates. This has allowed the haplotype to be

maintained for a prolonged period. If the introgressed segment had not fallen in a region with such a strong reduction in recombination rate, we would likely never have detected the signal of introgression as introgression fragments in other systems degenerate quickly, owing to recombination. The results of this study suggest that introgression between very divergent species might be more important in understanding evolutionary adaptation than previously thought.

URLs. dbSNP database, http://www.ncbi.nlm.nih.gov/projects/SNP/; NCBI Gene database, http://www.ncbi.nlm.nih.gov/gene.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The sequence reads are publicly available at the NCBI Sequence Read Archive under accession SRA096093.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank L. Andersson for critical discussions and reading of the manuscript. This study is supported by the National Key Research Project of China (2013ZX08006-5), the Natural Science Foundation of China (31230069) and the Changjiang Scholars and Innovative Research Team in University (IRT1136).

AUTHOR CONTRIBUTIONS

L. Huang and J.R. designed the study and analyzed the data. J.R., B.Y., H.A., X.F., R.N. and L. Huang wrote the manuscript. H.A., X.F., B.Y., Z.H., H.C., L.M., F.Z., L. Zhang, L.C., W.H., T.H., W.D. and R.N. performed bioinformatics analyses. J.Y., X.Y., L. Zhou, L. Han, J.L., S.S., X.X., B.L., Y.S., Y.L. and H.Y. collected data and performed sequencing and genotyping experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Larson, G. *et al.* Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* **307**, 1618–1621 (2005).
2. Larson, G. *et al.* Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. *Proc. Natl. Acad. Sci. USA* **107**, 7686–7691 (2010).
3. Wang, L. *et al.* in *Animal Genetic Resources in China: Pigs* (ed. China National Commission of Animal Genetic Resources) 2–16 (China Agricultural Press, 2011).
4. Nachman, M.W., Hoekstra, H.E. & D'Agostino, S.L. The genetic basis of adaptive melanism in pocket mice. *Proc. Natl. Acad. Sci. USA* **100**, 5268–5273 (2003).
5. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
6. Lamason, R.L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310**, 1782–1786 (2005).
7. Jones, F.C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
8. Kamberov, Y.G. *et al.* Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152**, 691–702 (2013).
9. Rubin, C.J. *et al.* Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587–591 (2010).
10. Atanur, S.S. *et al.* Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. *Cell* **154**, 691–703 (2013).
11. Axelsson, E. *et al.* The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**, 360–364 (2013).
12. Shapiro, M.D. *et al.* Genomic diversity and evolution of the head crest in the rock pigeon. *Science* **339**, 1063–1067 (2013).
13. Groenen, M.A. *et al.* Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393–398 (2012).
14. Rubin, C.J. *et al.* Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. USA* **109**, 19529–19536 (2012).
15. Li, M. *et al.* Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.* **45**, 1431–1438 (2013).
16. Fang, X. *et al.* The sequence and analysis of a Chinese pig genome. *Gigascience* **1**, 16 (2012).
17. Bosse, M. *et al.* Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS Genet.* **8**, e1003100 (2012).
18. Tortereau, 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).
19. Cornforth, M.N. & Eberle, R.L. Termini of human chromosomes display elevated rates of mitotic recombination. *Mutagenesis* **16**, 85–89 (2001).
20. Muotri, A.R., Marchetto, M.C., Coufal, N.G. & Gage, F.H. The necessary junk: new functions for transposable elements. *Hum. Mol. Genet.* 16 Spec. No. 2, R159–R167 (2007).
21. Boulant, J.A. & Dean, J.B. Temperature receptors in the central nervous system. *Annu. Rev. Physiol.* **48**, 639–654 (1986).
22. Adolph, E.F. & Molnar, G.W. Exchanges of heat and tolerances to cold in men exposed to outdoor weather. *Am. J. Physiol.* **146**, 507–537 (1946).
23. Chaffee, R.R. *et al.* Comparative chemical thermoregulation in cold- and heat-acclimated rodents, insectivores, protoprimates, and primates. *Fed. Proc.* **28**, 1029–1034 (1969).
24. Stocks, J.M., Taylor, N.A., Tipton, M.J. & Greenleaf, J.E. Human physiological responses to cold exposure. *Aviat. Space Environ. Med.* **75**, 444–457 (2004).
25. Jansky, L. & Hart, J.S. Cardiac output and organ blood flow in warm- and cold-acclimated rats exposed to cold. *Can. J. Physiol. Pharmacol.* **46**, 653–659 (1968).
26. Alazami, A.M. *et al.* Mutations in *C2orf37*, encoding a nucleolar protein, cause hypogonadism, alopecia, diabetes mellitus, mental retardation, and extrapyramidal syndrome. *Am. J. Hum. Genet.* **83**, 684–691 (2008).
27. Schmidt, E.M. *et al.* Chorein sensitivity of cytoskeletal organization and degranulation of platelets. *FASEB J.* **27**, 2799–2806 (2013).
28. Keatinge, W.R. *et al.* Increased platelet and red cell counts, blood viscosity, and plasma cholesterol levels during heat stress, and mortality from coronary and cerebral thrombosis. *Am. J. Med.* **81**, 795–800 (1986).
29. Ma, J. *et al.* Recombinational landscape of porcine X chromosome and individual variation in female meiotic recombination associated with haplotypes of Chinese pigs. *BMC Genomics* **11**, 159 (2010).
30. Nachman, M.W. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **17**, 481–485 (2001).
31. Choo, K.H. Why is the centromere so cold? *Genome Res.* **8**, 81–82 (1998).
32. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
33. Hewitt, G. The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913 (2000).
34. Frantz, L.A. *et al.* Genome sequencing reveals fine scale diversification and reticulation history during speciation in *Sus*. *Genome Biol.* **14**, R107 (2013).

ONLINE METHODS

Samples. A broad panel of 69 Chinese pigs from 11 diverse breeds and 3 wild boar populations were sequenced in this study (**Supplementary Table 1**). The 69 pigs were selected from 520 unrelated pigs (no common ancestor within 3 generations) pertaining to 32 Chinese breeds and 21 Chinese wild boar populations, which were previously genotyped for ~62,000 SNPs using porcine 60K DNA chips (Illumina)[35]. To select representative samples, individuals from each breed were chosen according to their genetic relationships in the neighbor-joining tree (**Supplementary Fig. 1**). The 69 pigs selected are highly representative of populations at the geographical extremes of China (**Fig. 1a**) and show good adaptation to high- or low-latitude environments (**Supplementary Table 1**). All experiments with pigs were performed under the guidance of ethical regulation from Jiangxi Agricultural University, China.

Genome sequencing. Genomic DNA was extracted from ear tissues using a standard phenol-chloroform method. For each individual, 1–15 μg of DNA was sheared into fragments of 200–800 bp using the Covaris system (Life Technologies). DNA fragments were then treated according to the Illumina DNA sample preparation protocol. Fragments were end repaired, A-tailed, ligated to paired-end adaptors and PCR amplified with 500-bp inserts for library construction. Sequencing was performed to generate 100-bp paired-end reads on the HiSeq 2000 platform (Illumina) according to the manufacture's standard protocols.

SNP calling. Filtered reads from all individuals were aligned to the Wuzhishan reference genome by the Burrows-Wheeler Aligner (BWA)[36]. To obtain high-quality SNPs, both SOAPsnp[37] and the Genome Analysis Toolkit (GATK)[38] were used for SNP calling for each sample. Population-based genotypes were created by combining sites for high-quality SNPs called in individuals. Before SNP calling, SAMtools[39] was used to sorting, merging and removing potential PCR duplications. Switches '-u' and '-t' were turned on to enable the rank-sum test for better accuracy, and the transition/transversion prior ratio was set to 2:1. Option '-m' was used for the sex chromosomes of male samples. SNPs with a read depth of less than 6 or greater than 70 or with a distance of less than 5 bp to a neighboring SNP were removed. For GATK SNP calling, standard preprocessing (including realignment and recalibration) and calling procedures were used as previously described[38]. SOAPsnp generated consensus genotypes for all genomic loci, even in regions with poor coverage. To remove low-quality genotypes called by SOAPsnp, we further excluded all loci that had a SOAPsnp quality score of less than 20 in any 1 of the 69 pigs sequenced. We identified 42.1 and 52.8 million SNPs by SOAPsnp and GATK, respectively. A common set of 40.8 million SNPs called by both GATK and SOAPsnp was chosen as the final 69-genome SNP data set. We also downloaded the publicly available whole-genome sequence data of 42 pigs (**Supplementary Table 7**)[13]. These 42 genomes were combined with the 69 genomes to create a 111-sample SNP data set. Population-based genotypes for the 111 pigs were created by GATK after BWA alignment and GATK preprocessing as described above. To avoid potential bias between our data and the publicly available data, we called SNPs by comparing the genome sequence of each individual to the Wuzhishan reference genome and then merged the called SNPs to form a common set of SNP data for the 111 individuals.

Structural variants. Indels were called along with SNPs by GATK for each sample. We used the in-house pipeline SeekSV v1.1.1 to identify structural variations for each sample. The underlying idea for the pipeline is similar to that for CREST[40]. Candidate structural variations with at least ten pairs of abnormal supporting reads were retained. Deletions were removed if their normalized coverage was greater than 5-fold. Most (88%) of the structural variations that we identified were deletions, which can be detected more effectively than other types of structural variation.

Loss-of-function variants. We defined variants as potential loss-of-function mutations if they corresponded to one of the following variants: (i) a SNP within a coding region resulting in a premature stop codons; (ii) a small indel within a coding region causing a frameshift of the ORF; (iii) a SNP or small indel within 2 bp of a splice site; and (iv) a structural variation overlapping a coding region. All loss-of-function variants were called using Perl scripts.

Validation of called SNPs. To check the accuracy of our sequence SNPs calling, we compared the SNP calls for the whole-genome sequencing data and the 60K data for the 69 sequenced pigs[35]. First, tagging sequences containing SNPs from the Illumina chip were mapped to the reference genome using BWA[36]. All mapped SNPs from the chip were further filtered by the same criteria for sequence-based SNP calling; those with a quality score less than 20 were removed.

Population genetics analysis. A neighbor-joining tree (**Fig. 1b**) was constructed with MEGA[41]on the basis of the IBS distance matrix data of all individuals calculated by PLINK v.107 (ref. 42). Identical scores (ISs) were calculated to evaluate the similarities of the sequenced genomes to the Wuzhishan reference genome according to the following formula:

$$\mathrm{IS} = \frac{\sum_{i=1}^{n} S_i}{2(n - n')}$$

where S_i is the number of alleles identical to the Wuzhishan reference allele at a given SNP site i, n is the total number of SNPs within a 50-kb window and n' is the total number of missing SNPs within a 50-kb windows.

Selective sweep analysis. LSBL statistics[43] were calculated for each polymorphic site with MAF >0.01 in the 69-genome data set on the basis of the fixation index (F_{ST}) values between three contrasting groups. Group A included Bamaxiang, Luchuan and Wuzhishan pigs from southern China, group B comprised Laiwu, Hetao and Min pigs from northern China and group C included Tibetan pigs from four geographical populations. Pairwise F_{ST} distances among the three groups (d_{AB}, d_{BC} and d_{AC}) were calculated according to Reich's formula[44]. We calculated x statistics for each SNP using the formula $x = (d_{AB} + d_{AC} - d_{BC})/2$. We performed the LSBL analysis in each 50-kb sliding window, with a step size of 25 kb. LSBL outliers, i.e., candidate selection loci, were defined as sites with x statistics surpassing the significance threshold ($P < 0.01$) determined using 10,000 permutation tests. A value of 0.618 was determined to be the threshold for LSBL outliers, corresponding to the top 0.35% of the empirical distributions of all tested SNPs. A candidate selective sweep was defined as a region containing at least five LSBL outlier SNPs within a 50-kb window. More than 97% of the LSBL outlier SNPs fell in the selective sweep regions we defined. Adjacent sweeps within a distance of 50 kb were merged into one sweep. Candidate genes under selection were defined as those overlapped by sweep regions or within 100 kb of LSBL SNP outliers. Two additional statistics, the absolute allele frequency difference between northern and southern Chinese pigs (ΔAF)[45] and standardized heterozygosity (Z_H) were calculated to confirm the signals in the top LSBL-defined sweep regions. Heterozygosity at each SNP was computed as the ratio of the heterozygous individuals to all individuals in a particular population. The average heterozygosity ($H_{50\ kb}$) was then calculated in each 50-kb sliding window, with a step size of 25 kb. Z_h is the standardized $H_{50\ kb}$ value over the whole genome. The haplotypes in target regions were inferred using the SHAPEIT2 program[46]. GO enrichment analysis for the annotated genes in the putative sweep regions was performed using the Cytoscape plugin program ClueGO[47], in which P values were corrected by the Benjamini-Hochberg approach.

Recombination rate calculation. All pigs from three F_2 intercross populations of White Duroc × Erhualian, Erhualian × Shaziling, and Erhualian × Bamaxiang were genotyped for ~62,000 SNPs on porcine 60K chips (Illumina) as described previously[48]. SNPs were filtered by excluding those with a MAF of <0.05, those with a call rate of <95% and those that were polymorphic in males at the X-Y specific region. Haplotype phasing was conducted with Phasebook software[49]. To identify crossover events that occurred in the germ line of the F_1 individuals, we exploited mendelian rules and linkage information to phase genotypes in the three F_2 populations. Crossover events were then identified as phase switches in the gametes transmitted by the F_1 parents to their offspring. Double crossovers occurring at intervals that were separated by less than three informative intervals were ignored.

doi:10.1038/ng.3199

Introgression analysis. Pairwise nucleotide differences per site within (d_x) and between (d_{xy}) populations were calculated by the following formulas:

$$d_x = \frac{2}{n_x(n_x-1)l}\sum_{i=1}^{n_x-1}\sum_{i'=i+1}^{n_x-1}k_{ii'}$$

$$d_{xy} = \frac{2}{n_x n_y l}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}k_{ij}$$

where k represents the number of differences between a particular pair of haplotypes in a target region. The subscripts i and j denote haplotypes from populations x and y, respectively, with primes indicating additional haplotypes from the same population. The expression for d_y is identical to that for d_x but with i replaced by j and n_x replaced by n_y. l denotes the effective length of the sequence without gaps in the target region. Divergence time was estimated by a calibrated molecular clock model $T = d/2r$, where r is the mutation rate of the number of genetic changes expected per unit time; d is the genetic distance of haplotypes between populations x and y, which was calculated using the formula $d = d_{xy} - (d_x + d_y)/2$; and T is the divergence time between populations x and y. First, we hypothesized that the time of divergence between *P. africanus* and *S. scrofa* was 9.9 million years ago, according to a recent report[34]. After we obtained d values using the above formula, we then calculated the parameter r. Finally, we estimated the divergence time between different populations using the calibrated mutation rate r. We calculated t_{MRCA} estimates using the method proposed by Thomson[50]. The mathematical formula is as follows:

$$\hat{T} = \sum_{i=1}^{n} x_i/(n\mu)$$

where $\hat{T}$ is the estimation of t_{MRCA}, x_i is the number of mutational differences between the ith sequence and the most recent common ancestor (MRCA), n is the total number of sequences in the sample and μ is the mutation rate. The sequence of the MRCA was inferred by maximum-likelihood phylogenetic tree using PHYLIP 3.69. Phylogenetic trees for the introgression and non-introgression regions were constructed using the Tamura-Nei model in MEGA[41].

35. Ai, H. *et al.* Inference of population history and genome-wide detection of signatures for high-altitude adaptation in Tibetan pigs. *BMC Genomics* **15**, 834 (2014).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
37. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
38. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
40. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8**, 652–654 (2011).
41. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
42. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
43. Shriver, M.D. *et al.* The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum. Genomics* **1**, 274–286 (2004).
44. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
45. Carneiro, M. *et al.* Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* **345**, 1074–1079 (2014).
46. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
47. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
48. Fan, Y. *et al.* A further look at porcine chromosome 7 reveals *VRTN* variants associated with vertebral number in Chinese and Western pigs. *PLoS ONE* **8**, e62534 (2013).
49. Druet, T. & Georges, M. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* **184**, 789–798 (2010).
50. Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J. & Feldman, M.W. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl. Acad. Sci. USA* **97**, 7360–7365 (2000).