SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
GENOME EVOLUTION
AGRICULTURAL GENETICS

Received
23 January 2014

Accepted
28 March 2014

Published
14 April 2014

Correspondence and requests for materials should be addressed to M.Z.L. (mingzhou.li@sicau.edu.cn); X.W.L. (xuewei.li@sicau.edu.cn) or R.Q.L. (lirq@pku.edu.cn)

* These authors contributed equally to this work.

Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication

Mingzhou Li[1,2]*, Shilin Tian[3]*, Carol K. L. Yeung[3]*, Xuehong Meng[3], Qianzi Tang[2], Lili Niu[2], Xun Wang[2], Long Jin[2], Jideng Ma[2], Keren Long[2], Chaowei Zhou[2,4], Yinchuan Cao[3], Li Zhu[2], Lin Bai[2], Guoqing Tang[2], Yiren Gu[5], An'an Jiang[2], Xuewei Li[2] & Ruiqiang Li[1,3]

[1]Biodynamic Optical Imaging Center (BIOPIC), Peking-Tsinghua Center for Life Sciences, and School of Life Sciences, Peking University, Beijing 100871, People's Republic of China, [2]Institute of Animal Genetics and Breeding, College of Animal Science and Technology, Sichuan Agricultural University, Ya'an 625014, People's Republic of China, [3]Novogene Bioinformatics Institute, Beijing 100083, People's Republic of China, [4]Department of Animal Science, Southwest University at Rongchang, Chongqing 402460, People's Republic of China, [5]Sichuan Animal Science Academy, Chengdu 610066, People's Republic of China.

Domesticated organisms have experienced strong selective pressures directed at genes or genomic regions controlling traits of biological, agricultural or medical importance. The genome of native and domesticated pigs provide a unique opportunity for tracing the history of domestication and identifying signatures of artificial selection. Here we used whole-genome sequencing to explore the genetic relationships among the European native pig Berkshire and breeds that are distributed worldwide, and to identify genomic footprints left by selection during the domestication of Berkshire. Numerous nonsynonymous SNPs-containing genes fall into olfactory-related categories, which are part of a rapidly evolving superfamily in the mammalian genome. Phylogenetic analyses revealed a deep phylogenetic split between European and Asian pigs rather than between domestic and wild pigs. Admixture analysis exhibited higher portion of Chinese genetic material for the Berkshire pigs, which is consistent with the historical record regarding its origin. Selective sweep analyses revealed strong signatures of selection affecting genomic regions that harbor genes underlying economic traits such as disease resistance, pork yield, fertility, tameness and body length. These discoveries confirmed the history of origin of Berkshire pig by genome-wide analysis and illustrate how domestication has shaped the patterns of genetic variation.

The over 730 pig (*Sus scrofa*) breeds or lines worldwide have undergone natural and artificial selection in various environments and produced high levels of phenotypic diversity, constituting a valuable resource for investigating how selection affects the genome[1,2]. Genes or genomic regions under selection in domesticated organisms, such as the 'domestication genes' found in silkworm[3], chicken[4], dog[5], cattle[6], yak[7] and pig[8,9], can be directly implicated in genetic breeding programs and greatly increases the efficiency of producing novel and desirable phenotypes[10]. The economic and biomedical importance of the domestic pig has led to significant efforts to decode the pig genome, such as that of the domestic Duroc pig[1] and Tibetan wild boars[9].

As part of a continuous effort to comprehensively document the genetic basis of pig phenotypic diversity, here we present genomic analyses among the European native Berkshire pig (three individuals) and other 38 pigs and wild boars distributed worldwide. The Berkshire pig is a typical traditional European breed that has been under intensive artificial selection since the early 18th century in England for rapid and efficient accumulation of muscle and desirable pork qualities such as juiciness, flavor, tenderness, pink-hued and heavily marbled. Using the whole genome sequencing approach, we explored the genetic relationships among Berkshire and other pigs, and identified genetic components under selection that are likely the consequence of domestication of Berkshire pig.

Results

Sequencing, mapping, SNP and InDel calling. Sequencing of three female Berkshire pigs generated a total of 36.65 Gb of paired-end DNA sequence, of which 36.29 Gb (99.02%) high quality paired-end reads were mapped to the pig reference genome assembly (Sscrofa10.2) (Supplementary Table S1). Consequently, for each individual, ~78.75% of reads mapped to 79.39% of the reference genome assembly with 3.64-fold average depth

(Supplementary Table S1). In addition, we also downloaded the genome data of 38 individuals from across the world from the EMBL-EBI database[1,11], including 14 European domestic pigs from five breeds, 6 Asian domestic pigs from three breeds in China, 7 Asian wild boars from four locations, 6 European wild boars from four locations, 4 other species in the genus *Sus*, and an African warthog. The average depth for the compiled dataset is 6.46-fold, with average mapping rate of 95.17% and ~75.25% coverage of the reference genome assembly (Supplementary Table S2).

We performed single-nucleotide polymorphism (SNP) calling and identified 18.68 million (M) SNPs from 41 individuals (Supplementary Table S3). We then pooled the SNPs into three groups, including 5.25 M from the 23 domestic pigs, 5.47 M from the 13 wild boars, and 15.73 M from the four wild genus *Sus* and an African warthog (Supplementary Table S3). A small portion of (2.48 M of 18.68 M, or 13.28%) SNPs was shared among the three groups, indicative of substantial genomic differences among them.

We identified 3.65 M SNPs from three Berkshire pigs, of which 21,905 coding SNPs leading to 7,773 nonsynonymous nucleotide substitutions (7,713 missense, 44 stop gain and 16 stop loss) were detected in 3,978 genes (Table 1 and Supplementary Data S1). Top 1,000 genes containing the highest number of nonsynonymous SNPs (nsSNPs) were mainly over-represented in olfactory-related categories, such as 'olfactory transduction (52 genes, $P = 1.68 \times 10^{-10}$)', 'sensory perception of smell (53 genes, $P = 8.70 \times 10^{-9}$)', 'olfactory receptor activity (52 genes, $P = 1.19 \times 10^{-8}$)' and 'sensory perception (77 genes, $P = 2.35 \times 10^{-7}$)' (Supplementary Data S2). The olfactory receptors, known to be involved in sensing of the extracellular environment, are encoded by the largest gene superfamily in the mammalian genome[12,13]. Pigs have one of the largest repertoire of functional olfactory receptor genes[14], reflecting the strong reliance of pigs on their sense of smell while scavenging for food[1] and other odor-driven behavior (particularly mate recognition and sexual receptive behavior)[15,16].

We also identified 2.93 M small insertion or deletion polymorphisms (InDels) ranging from 1–30 bp in length (Supplementary Table S4), which tend to be detected with greater frequency than long InDels. Only 2,991 (0.10%) InDels were located in coding sequences, of which 29.35% were multiples of 3 bp (Supplementary Fig. S1 and Data S3). The enrichment of in-frame InDels that are expected to preserve reading frame, can be explained by previous findings that in-frame InDels were under weaker negative selection than frameshift InDels with lengths that are not evenly divided by three[17,18]. These InDels affected genes enriched mainly in terms related to basic cellular functions, such as the 'binding of adenyl nucleotide, purine nucleoside, ATP, cation, ion, metal ion, and nucleoside' and 'protein kinase activity' (Supplementary Data S4), which is similar to previous reports in mammals about the effect of InDels on the functions of genes[1,5,18,19].

Table 1 | Summary and annotation of SNPs in Berkshire pigs

Category		Number of SNPs
Total		3,645,294
Upstream		23,796
Exonic	Missense	7,713
	Stop gain	44
	Stop loss	16
	Synonymous	14,132
Intronic		792,471
Splicing		118
Downstream		23,597
Upstream/Downstream		243
Intergenic		2,783,164

The package ANNOVAR[56] was used to identify whether SNPs cause protein coding changes and amino acids that are affected. 'Upstream' refers to a variant that overlaps with the 1 kb region upstream of the gene start site. 'Stop gain' means that an nsSNP leads to the creation of a stop codon at the variant site. 'Stop loss' means that an nsSNP leads to the elimination of a stop codon at the variant site. 'Splicing' means that a variant is within 2 bp of a splice junction. 'Downstream' means that a variant overlaps with the 1 kb region downstream of the gene end site. 'Upstream/Downstream' means that a variant is located in downstream and upstream regions (possibly for two different genes).

Phylogenetic and admixture analysis. To explore relatedness among the Berkshire pig and other pigs distributed worldwide, we conducted principle component analysis (PCA) using genomic SNPs. The first eigenvector geographically distinguishes 23 individuals in Europe from 17 individuals in Asian and a warthog in Africa, whereas the second eigenvector captures the biological differentiation between pigs (including domestic and wild boar) and other outgroup (i.e., wild genus *Sus* and warthog) (Fig. 1a and Supplementary Table S5). The neighbor-joining (NJ) tree confirmed these findings and further revealed genetically distinct clusters that relate to geographic locations rather than by domestic versus wild (Fig. 1b). This result is consistent with a deep phylogenetic split between European and Asian pigs since domestication about 10,000 years ago in multiple locations across Eurasia[1,20].

It is well documented that a clear signal for admixture between domestic pigs in Asia and Europe[1,9,21] is likely due to the importation of Chinese breeds into Europe (especially UK) at the onset of the agricultural revolution in the late 18th and 19th century[22]. To investigate the amount genetic material of Chinese origin in the Berkshire pigs relative to that shared by other five reprehensive European domestic pigs, we performed an admixture analysis (*D*-statistics) using 'ABBA/BABA' single nucleotide sites, which was originally developed to test for admixture between Neanderthals and modern humans[23,24]. We divided the genome into N blocks and computed the variance of the statistics over the genome N times, leaving each block aside and derived a standard error using the theory of the Jackknife[24]. Given the standard error of the *D*-statistics of different block sizes were very similar, we used 2 Mb as the block size for further analyses (Supplementary Table S6).

The excess of ABBA sites ($0 < D < 1$) indicates that the Berkshire pig has a stronger signal of introgression from Chinese domestic pig than other 5 European domestic pigs (Fig. 1c). Especially, when compared with Duroc pig, the Berkshire pig exhibits an highest excess of ABBA sites across 18 autosomes, giving a significantly positive D of 0.337 ± 0.010 (two-tailed Z-test for $D = 0$, $P = 1.680 \times 10^{-243}$) (Fig. 1c). The higher amount of Chinese genetic material in the Berkshire pig is consistent its history of origin: in the county of Berkshire in England, a reddish or sandy colored pig strain (sometimes spotted) was latterly refined with a cross of Siamese and Chinese blood (~300 years ago), bringing the color pattern we see today along with more efficient meat production[22]. Currently, the purebred Berkshire is recorded as a 'transboundary' (occurring in more than one country) breed.

Genome-wide selective sweep signals. To accurately detect the genomic footprints left by selection, we measured the genome-wide variations between six European wild boars and three Berkshire pigs, which are geographically close and genetically indistinguishable.

Compared with the wild boars, the domestic Berkshire pigs have lower levels of linkage disequilibrium (LD) across the range of distances separating loci ($P < 10^{-16}$, Mann-Whitney U test) (Fig. 2a), reflecting relatively higher inbreeding under artificial breeding programs and thus a lower genomic diversity in Berkshire pig.

Out of 272,292 windows of 100 kb in length sliding in 10 kb steps across the pig genome, 210,266 windows contain ≥ 50 SNP and cover 77.22% of the genome (Supplementary Fig. S2), which were used to detect signatures of selective sweeps. We used an empirical procedure and selected windows simultaneously with significantly high $\log_2$ (θ_π ratio ($\theta_{\pi,\text{ wild boar}}/\theta_{\pi,\text{ Berkshire}}$)) (10% right tail, where $\log_2$ (θ_π ratio) is 3.14) and significantly high F_{ST} values (10% right tail,

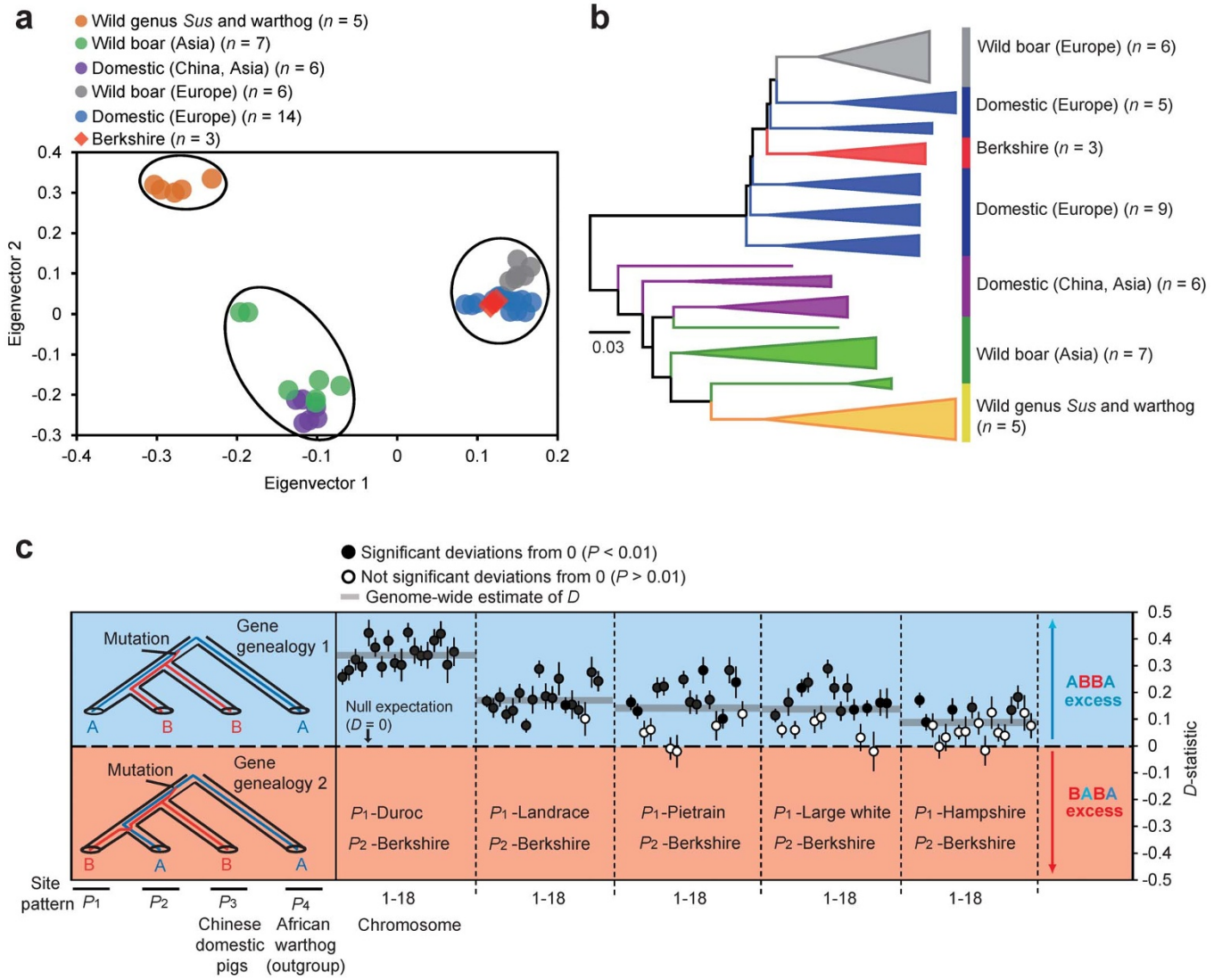

Figure 1 | Phylogenetic relationship and gene introgression. (a) Two-way PCA plot of pig breeds. The fraction of the variance explained is 33.56% for eigenvector 1 and 9.56% for eigenvector 2 with a Tracy-Widom *P* value $< 10^{-6}$ (Supplementary Table S5). (b) NJ phylogenetic tree of pig breeds. The scale bar represents *p* distance. (c) Four-taxon ABBA/BABA test of introgression. First panel from the left: ABBA and BABA nucleotide sites employed in the test are derived (- - *B* -) in Chinese domestic pigs compared with the warthog outgroup (- - - *A*), but differ among Berkshire and other 5 European domestic pigs (either ABBA or BABA). As this almost exclusively restricts attention to sites polymorphic in the ancestor of Chinese domestic pigs, Berkshire and other 5 European domestic pigs, equal numbers of ABBA and BABA sites are expected under a null hypothesis of no introgression, as depicted in the two gene genealogies. Second to last panel from the left: Distribution among chromosomes of *D*-statistic (± s.e.), which measures excess of ABBA sites over BABA sites, here for the comparison: Other 5 European domestic pigs (i.e. Duroc, Landrace, Pietrain, Large white and Hampshire), Berkshire, Chinese domestic pigs, African warthog.

where F_{ST} is 0.71) of the empirical distribution as regions with strong selective sweep signals along the genome, which should harbor genes that underwent selective sweep. Consequently, we identified a total of 11.95 Mb genomic regions (4.75% of the genome, containing 482 genes) with strong selective sweep signals in Berkshire pigs (Fig. 2b), which also exhibited significant differences ($P < 10^{-16}$, Mann-Whitney *U* test) in $\log_2 \theta_\pi$ ratio and F_{ST} values when compared to genomic background (Fig. 2c). SNPs from these regions formed two distinct clusters (i.e. Berkshire pigs and European wild boars) (Supplementary Fig. S3).

In total, 482 genes embedded in selected regions were predominantly related to immune (such as 'defense response to virus (4 genes, $P = 0.001$)' and 'Immunoglobulin' (11 genes, $P = 0.001$)), growth (such as 'regulation of growth (11 genes, $P = 0.003$)'), reproduction (such as 'oocyte meiosis (6 genes, $P = 0.004$)' and 'reproductive developmental process (9 genes, $P = 0.005$)') (Table 2). This result coincides with previous reports of pig domestic genes[1,8,11,21,25] and may be responsible for dramatic phenotypic changes in domestic pigs that are of economic values, such as disease resistance, pork yield and fertility. In addition, we also identified genes related to neuron functions (such as 'neurotrophin signaling pathway (8 genes, $P = 0.003$)') that experienced selective sweep (Table 2), which support the hypothesis that selection for altered behavior (such as tameness or aggression towards humans) was important during pig domestication and that mutations affecting developmental genes may underlie these changes[10,26,27]. For example, one of the genes under selective sweep in Berkshire pig is transcription factor *SOX6* (SRY (sex determining region Y)-box 6) (Supplementary Data S5), a modulator of cell fate during neocortex development[28], which plays roles in brain development and related to the differences in the development or maturation of the frontal cortex in domesticated animals[29].

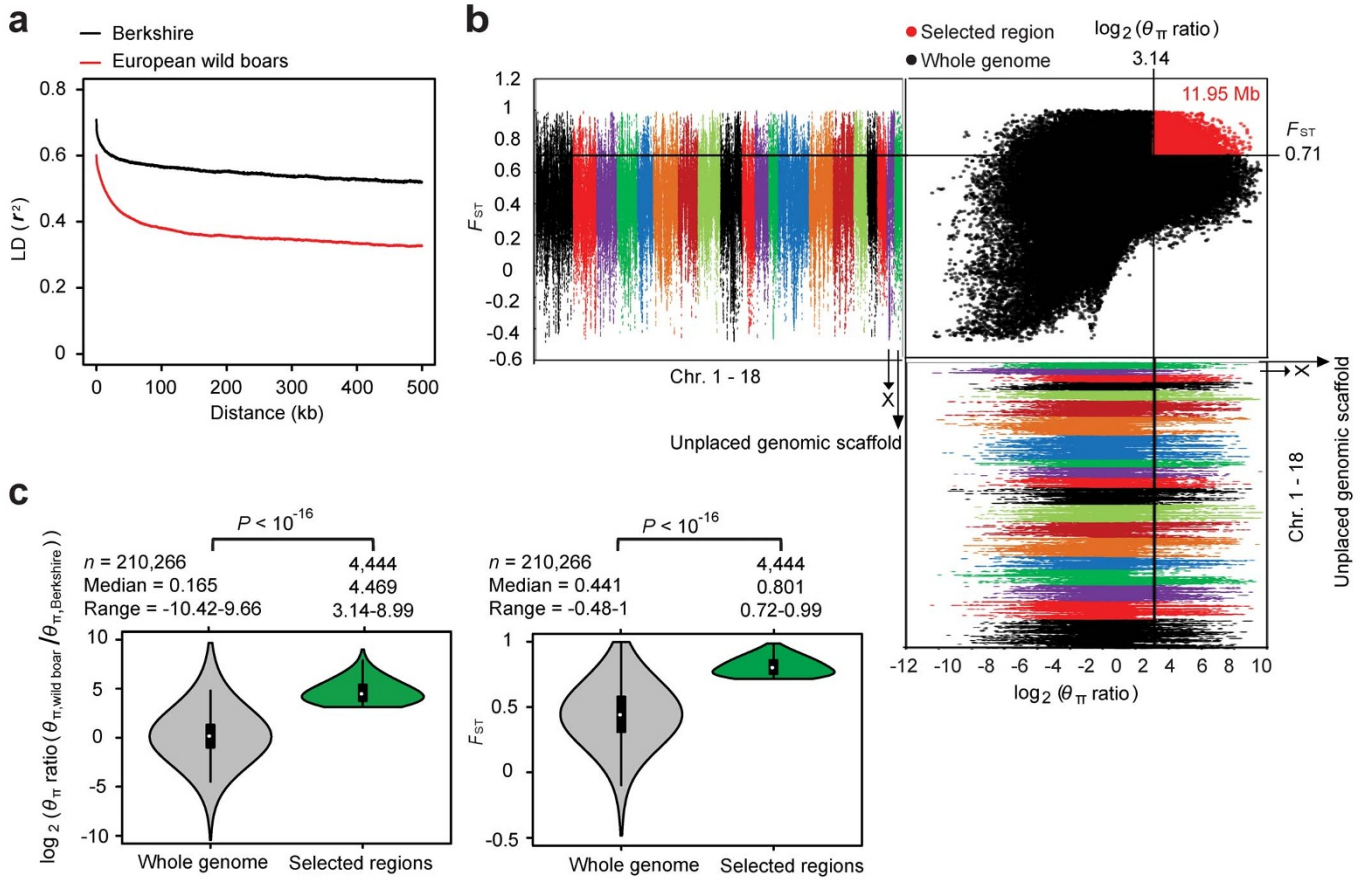

Figure 2 | Identification of genomic regions with strong selective sweep signals in Berkshire pigs. (a) LD patterns of Berkshire and European wild boars. (b) Distribution of $\log_2$ (θ_π ratio ($\theta_{\pi,\ \text{wild boar}}/\theta_{\pi,\ \text{Berkshire}}$)) and F_{ST}, which are calculated in 100 kb windows sliding in 10 kb steps. Data points located to the right of the vertical lines (corresponding to 10% right tails of the empirical $\log_2$ (θ_π ratio) distribution, where $\log_2$ (θ_π ratio) is 3.14) and above the horizontal line (10% right tail of the empirical F_{ST} distribution, where F_{ST} is 0.71) were identified as selected regions for Berkshire pigs (red points). (c) Violin plot of θ_π ratio and F_{ST} values for regions of Berkshire pigs that have undergone positive selection versus the whole genome. Each "violin" with the width depicting a 90°-rotated kernel density trace and its reflection. Vertical black boxes denote the interquartile range (IQR) between the first and third quartiles (25th and 75th percentiles, respectively) and the white point inside denotes the median. Vertical black lines denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. The statistical significance was calculated by the Mann-Whitney *U* test.

Body length in domestic pigs. Notably, we detected numerous well-characterized genes related to body length embedded in selected regions (Fig. 3a), which is the most characteristic morphological change between the wild boar and domestic pig. Wild boars, which are ancestors of domestic pigs, have 19 vertebrae. In comparison, European commercial breeds have 21–23 vertebrae, probably owing to selective breeding for enlargement of body size[30].

Eight genes exhibiting strong selective sweep signals are significantly over-represented in 'OMIM-disease term: Many sequence variants affecting diversity of adult human height' ($P = 0.002$)' (Supplementary Data S5), which has been documented to associate significantly with adult human height[31]. For example, *ADAMTSL3* (a disintegrin-like and metalloprotease domain with thrombospondin type I motifs-like 3), a glycoprotein in extracellular matrix, is

Table 2 | Top ten functional gene categories enriched for genes affected by domestication

Category	Term description	Involved gene number	*P* value
GO-BP: 0051607	Defense response to virus	4	0.001
InterPro: 013151	Immunoglobulin	11	0.001
KEGG-pathway: 04722	Neurotrophin signaling pathway	8	0.003
GO-BP: 0040008	Regulation of growth	11	0.003
InterPro:007110	Immunoglobulin-like	16	0.004
KEGG-pathway:04114	Oocyte meiosis	6	0.004
GO-BP:0009615	Response to virus	6	0.005
GO-BP:0003006	Reproductive developmental process	9	0.005
GO-BP:0045137	Development of primary sexual characteristics	6	0.005
GO-MF: 0005267	Potassium channel activity	8	0.013

P values (i.e. EASE scores), indicating significance of the overlap between various gene sets, were calculated using a Benjamini-corrected modified Fisher's exact test. A complete list of categories and gene names are provided in Supplementary Data S5.

associated with the chondrogenesis, morphogenesis and growth of the skeleton in human[31–33] and other mammals (cattle)[34], which is also an attractive candidate genetic marker to identify animal body size or type. *GPR126* (G-protein coupled receptor 126), an orphan receptor of the adhesion-G-protein coupled receptor family, is essential for mammalian embryonic viability[35], myelination[36], osteoclast function and regulation of bone mineral density[37]. Association between variation at *GPR126* with height in childhood[38] and adult[31,33] as well as the skeletal frame size[39] has been shown. *PRKG2* (cGMP-dependant type II protein kinase) is involved in preovulatory follicles as a response to luteinizing hormone and progesterone, which is highly expressed in brain and in cartilage, and contributed to the determination of dwarfism in mammals. The knockout mouse[40] and naturally occurring rat[41] and cattle[42] *PRKG2* mutants resulted in unorganized growth plate with abnormal stacking of chondrocytes and dwarfism. In addition, the *RASGEF1B* (RasGEF domain family, member 1B) is a highly conserved guanine nucleotide exchange factor for Ras family proteins[43], which is neighboring with the *PRKG2*. Ras superfamily proteins function as molecular switches in fundamental events such as signal transduction, cytoskeleton dynamics and intracellular trafficking. In human, the microdeletion (1.37 Mb) at chromosome 4q21 that encompass *PRKG2* and *RASGEF1B* resulted in growth restriction, mental retardation and absent or severely delayed speech[43].

We also found *IGF1* (Insulin-like growth factor 1), a hormone similar in molecular structure to insulin, which is a primary mediator of the effects of growth hormone, could stimulate systemic body growth, especially skeletal muscle, cartilage and bone, and has been recognized as a major determinant of body size in mammals[44,45]. In particular, *NR6A1* (nuclear receptor subfamily 6, group A, member 1), which is involved in neurogenesis and germ cell development[46], to be embedded in the most significantly selected regions (simultaneously with high $\log_2$ (θ_π ratio) (1% right tail) and F_{ST} values (1% right tail) (Fig. 3b). It has been well documented that the *NR6A1* is a strong candidate for being a causal gene underlying the elongation of the back and an increased number of vertebrae in pigs varies[8,47,48].

Analogously with the human height[31], the porcine trunk length is a highly heritable (the number of vertebrae, $h^2 = 0.62$) and classic polygenic trait[49]. The strong selective sweeps of these genes related to 'body length' reveal the specific evolutionary scenarios triggered by artificial selection for agricultural production.

Conclusions

This study presented the genetic relationships between the Berkshire and other pigs, and uncovered genetic footprints of domestication that provide an important resource for further improvements of this important livestock species. The work performed here will serve as a typical demonstration for future deciphering the genomic differences shaped by the artificial selections.

Methods

Ethics statement. All research involving animals were conducted according to the Regulations for the Administration of Affairs Concerning Experimental Animals (Ministry of Science and Technology, China, revised in June 2004) and approved by the Institutional Animal Care and Use Committee in College of Animal Science and

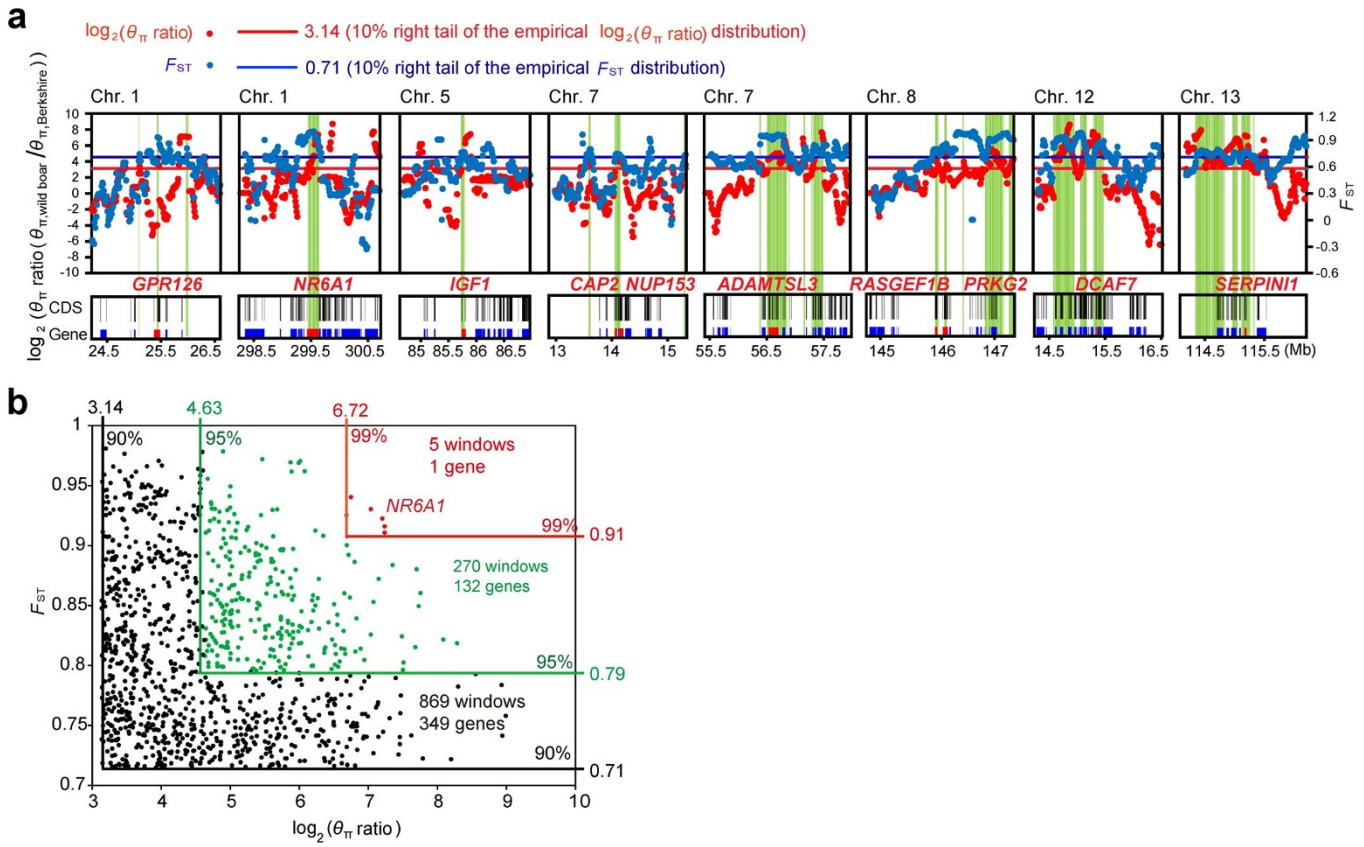

Figure 3 | Genes related to body length with strong selective sweep signals in Berkshire pigs. (a) Log_2 (θ_π ratio ($\theta_{\pi,\text{ wild boar}}/\theta_{\pi,\text{ Berkshire}}$)) and F_{ST} values are plotted using a 10 kb sliding window for genes embedded in selected regions. Genomic regions located above the upper horizontal blue line (corresponding to a 10% significance level of F_{ST}, where $F_{ST} = 0.71$) and above the lower horizontal red line (a 10% significance level of θ_π ratio, where $\log_2$ (θ_π ratio) $= 3.14$) were termed as regions with strong selective sweep signals (green regions). Genome annotations are shown at the bottom (black bar: coding sequences, blue bar: genes). The boundary of ten genes related to body length is marked in red. (b) *NR6A1* gene with strong selective sweep signals. Out of 482 genes embedded in selected regions which crossed 1,144 windows of 100 kb in length sliding in 10 kb steps, only one gene (i.e. *NR6A1*) is embedded in the most significantly (1% right tail $\log_2$ (θ_π ratio and F_{ST} values) selected regions ($\log_2$ (θ_π ratio) $= 6.72$; $F_{ST} = 0.91$).

Technology, Sichuan Agricultural University, Sichuan, China under permit No. DKY- S20123130.

Sequencing of Berkshire pigs. Genomic DNA was extracted from the ear tissues of each of three female Berkshire pigs. There is no direct and collateral blood relationship within the last three generations among them. Sequencing was performed on the Illumina HiSeq 2000 platform. In addition, we also downloaded the genome data of 38 *Sus scrofa* individuals across the world from the EMBL-EBI database (ftp.sra.ebi.ac.uk/vol1/fastq/ERR173/).

Sequence quality checking and filtering. First, to avoid reads with artificial bias (i.e. low quality paired reads, which mainly result from base-calling duplicates and adapter contamination), we removed the following types of reads: (a) reads with $\geq$ 10% unidentified nucleotides (N); (b) reads with $>$ 10 nt aligned to the adapter, allowing $\leq$ 10% mismatches; and (c) reads with $>$ 50% bases having phred quality $<$ 5; and (d) putative PCR duplicates generated by PCR amplification in the library construction process (i.e. read 1 and read 2 of two paired-end reads that were completely identical). Second, high quality paired-end reads were mapped to the pig reference genome sequence (Sscrofa10.2) using the BWA software[50]. The reference was indexed and the command 'aln -o 1 -e 10 -t 4 -l 32 -i 15 -q 10' was used to find the suffix array coordinates of good matches for each read. The best alignments were generated in the SAM format given paired-end reads with command 'sampe'. We further improved the alignment results with the following three steps: (a) filter the alignment read with mismatches $\leq$ 5 and mapping quality = 0; (b) the alignment results were corrected using the package Picard (http://sourceforge.net/projects/picard/) with two core commands. The 'AddOrReplaceReadGroups' command was used to replace all read groups in the INPUT file with a new read group and assigns all reads to this read group in the OUTPUT BAM. 'FixMateInformation' command was used to ensure that all mate-pair information was in sync between each read and its mate pair; and (c) remove potential PCR duplication. If multiple read pairs have identical external coordinates, only the pair with the highest mapping quality was retained.

SNP and InDel calling. After alignment, we performed SNP calling on a population-scale for three groups (23 domestic pigs, 13 wild boars, and four species of the wild genus *Sus* and an African warthog) using a Bayesian approach as implemented in the package SAMtools[51]. The genotype likelihoods from reads for each individual at each genomic location were calculated, and the allele frequencies were also estimated. The 'mpileup' command was used to identify SNPs with the parameters as '-q 1 -C 50 -S -D -m 2 -F 0.002 –u'. Then, to exclude SNP calling errors caused by incorrect mapping or InDels, only high quality SNPs (coverage depth $\geq$ 4 and $\leq$ 1,000, RMS mapping quality $\geq$ 20, the distance between adjacent SNPs $\geq$ 5 bp, no InDel present within a 3 bp window and the missing ratio of samples within each group $<$ 50%) were kept for subsequent analysis. We also performed InDel calling using the 'mpileup' command with the parameters as '-m 2 -F 0.002 -d 1,000' as implemented in the package SAMtools[51].

Functional enrichment analysis. Functional enrichment analysis of Gene Ontology (GO), pathway and InterPro domains was performed using the DAVID web server[52]. Genes were mapped to their respective human orthologs, and the lists were submitted to DAVID for enrichment analysis of the significant overrepresentation of GO biological processes (GO-BP), molecular function (GO-MF) terminologies, and KEGG-pathway and InterPro categories. In all tests, the whole known genes were appointed as the background, and *P* values (i.e. EASE score), indicating significance of the overlap between various gene sets, were calculated using Benjamini-corrected modified Fisher's exact test. Only terms with a *P* value less than 0.05 were considered as significant and listed.

Phylogenetic genetic analyses. We performed the PCA with the population scale SNPs using the package EIGENSOFT4.2[53], and the eigenvectors were obtained from the covariance matrix using the R function reigen. The significance level of eigenvectors was determined using the Tracey-Widom test[53]. The phylogenetic tree was inferred using TreeBeST (http://treesoft.sourceforge.net/treebest.shtml) under the *p*-distances model using SNPs in a population scale.

Admixture analysis – *D*-statistics (ABBA-BABA tests). To detect admixture between Chinese domestic pigs and Berkshire or other five European domestic pigs, we computed *D*-statistics based on ABBA and BABA SNP frequency differences using the expression[23]:

$$D(P_1, P_2, P_3, P_4) = \frac{\sum_{i=1}^{n} [(1-p_{i1})p_{i2}p_{i3}(1-p_{i4}) - p_{i1}(1-p_{i2})p_{i3}(1-p_{i4})]}{\sum_{i=1}^{n} [(1-p_{i1})p_{i2}p_{i3}(1-p_{i4}) + p_{i1}(1-p_{i2})p_{i3}(1-p_{i4})]} \quad (1)$$

where P_1, P_2, P_3 and P_4 are the four different populations under comparison, P_1 (separately, each of 5 European domestic pigs) and P_2 (Berkshire) are sister taxa, P_3 is the Chinese domestic pigs and P_4 (African warthog) is an outgroup, p_{ij} is the observed frequency of the derived "B" SNP i in taxon j, and n is the total number of SNPs.

It is possible to compute the number of derived alleles common between P_1 and P_3 (ABBA count) and between P_2 and P_3 (BABA count). Under the null hypothesis of solely incomplete lineage sorting and no gene flow between P_3 and either P_2 or P_1, we expect a similar count of ABBA and BABA patterns. Under an alternative scenario of gene flow, the count of ABBA must be significantly higher than BABA counts (or vice versa)[23,24,54]. In addition to calculating *D* for the entire genome, to examine variation in *D* across the genome, separate *D*-statistics were evaluated for each of the 18 autosomes. A standard error (s.e.) of the *D*-statistics was computed using a Weighted Block Jackknife approach[24].

Linkage-disequilibrium (LD) analysis. To estimate the LD patterns between Berkshire and European wild boars, we used 4.92 M SNPs of six European wild boars and merged them with SNPs of the Berkshire pigs, resulting in 8.57 M SNPs in total. To evaluate LD decay, the coefficient of determination (r^2) between any two loci was calculated using Haploview[55]. Average r^2 was calculated for pairwise markers in a 500 kb window and averaged across the whole genome.

Calculation of θ_π and F_{ST}. A sliding window approach (100 kb windows sliding in 10 kb steps) was applied to quantify the polymorphism levels (θ_π, pairwise nucleotide variation as a measure of variability) and genetic differentiation (F_{ST}) between the Berkshire and European wild boars.

1. Groenen, M. A. *et al.* Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393–398 (2012).
2. Chen, K., Baxter, T., Muir, W. M., Groenen, M. A. & Schook, L. B. Genetic resources, genome mapping and evolutionary genomics of the pig (*Sus scrofa*). *Int. J. Biol. Sci.* **3**, 153–165 (2007).
3. Xia, Q. *et al.* Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**, 433–436 (2009).
4. Rubin, C. J. *et al.* Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587–591 (2010).
5. Axelsson, E. *et al.* The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**, 360–364 (2013).
6. Gibbs, R. A. *et al.* Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**, 528–532 (2009).
7. Wang, K. *et al.* Genome-wide variation within and between wild and domestic yak. *Mol. Ecol. Resour.* **14**, 3; DOI: 10.1111/1755-0998.12226 (2014).
8. Rubin, C. J. *et al.* Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. USA* **109**, 19529–19536 (2012).
9. Li, M. *et al.* Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.* **45**, 1431–1438 (2013).
10. Larson, G. & Burger, J. A population genetics view of animal domestication. *Trends Genet.* **29**, 197–205 (2013).
11. Bosse, M. *et al.* Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS Genet.* **8**, e1003100 (2012).
12. Niimura, Y. Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. *Curr. Genomics* **13**, 103–114 (2012).
13. Plessy, C. *et al.* Promoter architecture of mouse olfactory receptor genes. *Genome Res.* **22**, 486–497 (2012).
14. Nguyen, D. T. *et al.* The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. *BMC Genomics* **13**, 584 (2012).
15. Baum, M. J. Contribution of pheromones processed by the main olfactory system to mate recognition in female mammals. *Front. Neuroanat.* **6**, 20 (2012).
16. Mak, G. K. *et al.* Male pheromone-stimulated neurogenesis in the adult female brain: possible role in mating behavior. *Nat. Neurosci.* **10**, 1003–1011 (2007).
17. Chen, F., Chen, C., Li, W. & Chuang, T. Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res.* **17**, 16–22 (2007).
18. Li, Y. *et al.* Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly. *Nat. Biotechnol.* **29**, 723–730. (2011).
19. Feuk, L. *et al.* Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genetics* **1**, e56. (2005).
20. Larson, G. *et al.* Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* **307**, 1618–1621 (2005).
21. Wilkinson, S. *et al.* Signatures of diversifying selection in European pig breeds. *PLoS Genet.* **9**, e1003453 (2013).
22. White, S. From globalized pig breeds to capitalist pigs: a study in animal cultures and evolutionary history. *Environ. Hist.* **16**, 94–120 (2011).
23. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
24. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
25. Amaral, A. J. *et al.* Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS ONE* **6**, e14782 (2011).
26. Albert, F. W. *et al.* Targeted resequencing of a genomic region influencing tameness and aggression reveals multiple signals of positive selection. *Heredity* **107**, 205–214 (2011).
27. Hare, B., Wobber, V. & Wrangham, R. The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. *Anim. Behav.* **83**, 573 (2012).
28. Azim, E., Jabaudon, D., Fame, R. M. & Macklis, J. D. *SOX6* controls dorsal progenitor identity and interneuron diversity during neocortical development. *Nat. Neurosci.* **12**, 1238–1247 (2009).

29. Albert, F. W. *et al.* A comparison of brain gene expression levels in domesticated and wild animals. *PLoS Genet.* **8**, e1002962 (2012).
30. King, J. & Roberts, R. Carcass length in the bacon pig: its association with vertebrae numbers and prediction from radiographs of the young pig. *Anim. Prod.* **2**, 59–65 (1960).
31. Gudbjartsson, D. F. *et al.* Many sequence variants affecting diversity of adult human height. *Nat. Genet.* **40**, 609–615 (2008).
32. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* **40**, 575–583 (2008).
33. Lettre, G. *et al.* Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.* **40**, 584–591 (2008).
34. Liu, Y. *et al.* Molecular characterization, expression pattern, polymorphism and association analysis of bovine *ADAMTSL3* gene. *Mol. Biol. Rep.* **39**, 1551–1560 (2012).
35. Waller-Evans, H. *et al.* The orphan adhesion-GPCR *GPR126* is required for embryonic development in the mouse. *PLoS ONE* **5**, e14047 (2010).
36. Monk, K. R., Oshima, K., Jors, S., Heller, S. & Talbot, W. S. *GPR126* is essential for peripheral nerve development and myelination in mammals. *Development* **138**, 2673–2680 (2011).
37. Hsiao, E. C. *et al.* Osteoblast expression of an engineered Gs-coupled receptor dramatically increases bone mass. *Proc. Natl. Acad. Sci. USA* **105**, 1209–1214 (2008).
38. Zhao, J. *et al.* The role of height-associated loci identified in genome wide association studies in the determination of pediatric stature. *BMC Med. Genet.* **11**, 96 (2010).
39. Soranzo, N. *et al.* Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. *PLoS Genet.* **5**, e1000445 (2009).
40. Pfeifer, A. *et al.* Intestinal secretory defects and dwarfism in mice lacking cGMP-dependent protein kinase II. *Science* **274**, 2082–2086 (1996).
41. Chikuda, H. *et al.* Cyclic GMP-dependent protein kinase II is a molecular switch from proliferation to hypertrophic differentiation of chondrocytes. *Gene. Dev.* **18**, 2418–2429 (2004).
42. Koltes, J. E. *et al.* A nonsense mutation in cGMP-dependent type II protein kinase (*PRKG2*) causes dwarfism in American Angus cattle. *Proc. Natl. Acad. Sci. USA* **106**, 19250–19255 (2009).
43. Bonnet, C. *et al.* Microdeletion at chromosome 4q21 defines a new emerging syndrome with marked growth restriction, mental retardation and absent or severely delayed speech. *J. Med. Genet.* **47**, 377–384 (2010).
44. Sutter, N. B. *et al.* A single *IGF1* allele is a major determinant of small size in dogs. *Science* **316**, 112–115 (2007).
45. Niu, P. *et al.* Porcine insulin-like growth factor 1 (*IGF1*) gene polymorphisms are associated with body size variation. *Genes & Genomics*, 1–6 (2013).
46. Zhao, H., Li, Z., Cooney, A. J. & Lan, Z. Orphan nuclear receptor function in the ovary. *Front. Biosci.* **12**, 3398–3405 (2007).
47. Mikawa, S. *et al.* Fine mapping of a swine quantitative trait locus for number of vertebrae and analysis of an orphan nuclear receptor, germ cell nuclear factor (*NR6A1*). *Genome Res.* **17**, 586–593 (2007).
48. Yang, G., Ren, J., Zhang, Z. & Huang, L. Genetic evidence for the introgression of Western *NR6A1* haplotype into Chinese Licha breed associated with increased vertebral number. *Anim. Genet.* **40**, 247–250 (2009).
49. Borchers, N., Reinsch, N. & Kalm, E. The number of ribs and vertebrae in a Pietrain cross: variation, heritability and effects on performance traits. *J. Anim. Breed. Genet.* **121**, 392–403 (2004).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
52. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
53. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
54. Dasmahapatra, K. K. *et al.* Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
55. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
56. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

Acknowledgments

This work was supported by grants from the National High Technology Research and Development Program of China (863 Program) (2013AA102502), the National Special Foundation for Transgenic Species of China (2014ZX0800950B and 2011ZX08006-003), the Fund of Fok Ying-Tung Education Foundation (141117), the Fund for Distinguished Young Scientists of Sichuan Province (2013JQ0013), the Specialized Research Fund of Ministry of Agriculture of China (NYCYTX-009), the Program for Changjiang Scholars and Innovative Research Team in University (IRT13083), the Postdoctoral Fellowship of Peking-Tsinghua Center for Life Sciences, and the China Postdoctoral Science Foundation (2012M520123).

Author contributions

M.L., S.T., C.K.L.Y., X.W. and R.L. led the experiments and designed the analytical strategy. L.N. and L.J. performed animal work and prepared biological samples. J.M., K.L. and C.Z. constructed the DNA library and performed sequencing. M.L., S.T., C.K.L.Y., X.M., X.W., Q.T., Y.C., L.Z., X.L. and G.T. designed the bioinformatics analysis process. M.L., S.T., Y.G., L.B. and A.J. wrote the paper. C.K.L.Y., X.L. and R.L. revised the paper.

Additional information

Accessions codes The genome resequencing reads have been deposited into the NCBI sequence read archive (SRA) under the accession SRP030626.

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

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

How to cite this article: Li, M.Z. *et al.* Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Sci. Rep.* **4**, 4678; DOI:10.1038/srep04678 (2014).