Li et al: Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple *De Novo* assemblies

Supplemental Methods

1 *De novo* sequencing and assembly of nine pig genomes

1.1 Animals

Five European and four Chinese pig breeds (a female individual at 60 days old for each breed) were selected, which have known history of origin and development and exhibited marked phenotypic diversity in appearance, fertility, growth, palatability and local fitness (**Fig. 2A** and **Supplemental Fig. S1**). Genomic DNA was extracted from the liver tissue for each of nine individuals using the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer's instructions.

1.2 Genome sequencing

Whole genome shotgun strategy and next-generation sequencing technologies on the Illumina HiSeq 2500 platform were used. Each genome was sequenced using a combination of short-insert (180 and 500 bp) and long-insert (2, 5, 6 and/or 10 kb) DNA libraries according to the manufacturer's specifications (Illumina), and read lengths were 100 bp (**Supplemental Fig. S2** and **Table S1**). After filtering out the adapter sequences (> 10 nt aligned to the adapter, allowing ≤ 10% mismatches), low quality reads (i.e. ≥ 10% unidentified nucleotides or > 50% bases having phred quality < 5) and duplicated reads, an average of 229.5 Gb high-quality data were retained for each assembly, of which the quality of 95.85% and 89.00% of the bases were ≥ Q20 and ≥ Q30, respectively (**Supplemental Table S1**).

1.3 Estimation of genome size

To estimate the genome size of the nine pigs, we selected an average of 143.84

Gb of high-quality short-insert reads (180 and 500 bp) for each breed, and generated 17-mer frequency information based on *K*-mer analysis as implemented in the software Meryl (Supplemental URLs) (**Supplemental Fig. S3**). We estimated the genome size of nine pigs to be in average of 2,297.78 Mb (~2.30 Gb) (**Supplemental Table S2**).

1.4 *De novo* assembly

Paired-end reads were processed using the error-correction module of ALLPATHS-LG[149] to remove base calling errors. Each genome was assembled using SOAPdenovo, a *de novo* genome assembler based on a *de Bruijn* graph algorithm[150,151]. We also used ErrorCorrection in SOAPdenovo package to connect 180 bp library paired-end reads and to generate longer sequences for assembly. Reads of 180 and 500 bp library were used for contig building, and all paired-end reads libraries were used to provide links for scaffold construction. GapCloser (v1.12) from SOAPdenovo package was used for gap filling within assembled scaffolds using all paired-end reads. As a result, the contig N50 sizes of nine assemblies ranged from 28.99 to 42.66 kb and scaffold N50 sizes ranged from 1.26 to 2.45 Mb. Consequently, a total length of the ungapped sequences of nine assemblies ranging from 2.45 to 2.49 Gb were obtained, which is similar to the amount generated for the reference assembly (2.52 Gb) (**Supplemental Figs. S4,5,6** and **Tables S3,4**). We also improved the Tibetan wild boar assembly[42] by increasing the contig N50 size from 20.69 kb to 22.54 kb and the ungapped genome assembly size from 2.43 Gb to 2.44 Gb (**Supplemental Table S4**).

2. Assessment of quality and completeness of assemblies

To validate the single-base accuracy of the genome assemblies, we realigned the high-quality reads of short-insert (180 and 500 bp) DNA libraries to the assemblies and found that an average of 95.86% of the reads could be mapped, and 97.48 % of the genome had coverage depth ≥ 10, indicating a high sequencing depth for the whole genomes (**Supplemental Fig. S7**).

To assess the completeness of the genome, we aligned the coding sequences (CDS) of the reference genome to the *de novo* assemblies, and mapped the high-quality reads of short-insert (180 and 500 bp) DNA libraries to

the CDS of the reference genome, which both yielded more than 99% coverage of the 21,566 protein-coding genes (Sscrofa10.2 annotation) in number and length by nucleotide sequence similarity (**Supplemental Table S5**). We further used the core eukaryotic genes mapping approach (CEGMA) pipeline (v.2.5)[44] to evaluate the nine pig genome assemblies. This effort identified 90.86% and 95.30% of 248 core eukaryotic genes (CEGs) completely or partially present in nine assemblies, respectively, which is comparable to that identified in Tibetan wild boar (Complete: 89.52%; Partial: 95.16%), cow (92.34%; 95.56%) and sheep (90.32%; 95.16%) genome assemblies and more than that identified in the reference pig genome assembly (82.26%; 91.53%) (**Supplemental Table S6**).

Together, the completeness of the assembled genomes enabled accurate detection of variation and comparative analyses within genic regions.

3. Repeat annotation

We performed repeat annotation for ten assemblies (**Supplemental Figs. S5,6**). For comparisons with the reference assembly, we also annotated repetitive elements in the reference pig genome using the same process and parameters.

(a) Identification of known transposable elements (TEs)

We used RepeatMasker (v.4.0.5) against the Repbase TE library (RM database, v.15.02)[152,153], and RepeatProteinMask (v.4.0.5) to perform WU-BLASTX against the TE protein database (Supplemental URLs).

(b) *De novo* repeat prediction

We built a *de novo* repeat library for the pigs using RepeatModeler (v.1.0.8,) which uses two core programs, i.e. RECON[154] and RepeatScout[155] to generate the predicted TE families (Supplemental URLs).

4. SNP calling using an assembly-versus-assembly method

To accurately identify SNPs between ten assemblies against the reference genome, we took advantage of an assembly-versus-assembly approach to identify the candidate SNP sites and further determined the heterozygous or homozygous SNPs by aligning short sequencing reads.

4.1 Gapped alignment of *de novo* assemblies to a reference genome

Assembled scaffolds of ten individuals were separately aligned to the reference genome using the LASTZ program (Supplemental URLs), with the parameters T = 2 (no transition), Y (ydrop) = 15,000, L (gappedthresh) = 3,000 and K (hspthresh) = 4,500. The raw alignments were combined into larger syntenic blocks using the ChainNet algorithm. The best hit of every single location on chromosomes was chosen by the utility “axtBest” based on a dynamic-programming algorithm, with the same substitution matrix adopted in alignment. The different base-types between the assembly and reference genome within the best alignment hits are the candidate SNP sites. Consequently, we identified 13.20 M such candidate SNP sites for each individual against the reference genome (**Fig. 1** and **Supplemental Table S8**).

4.2 Determination of the heterozygous or homozygous SNPs by aligning short sequencing reads

To further determine the genotype of SNPs in a diploid genome, we separately mapped the high quality paired-end short-insert reads (180 and 500 bp) of each individual (151.08 Gb) onto both independently assembled genomes (65.50-fold coverage per individual) and the reference genome (59.95-fold coverage per individual) using the BWA software (v.0.7.12)[43] with the command ‘mem -t 10 -k 32’. Then BAM alignment files were generated by package SAMtools (v.1.3)[3]. Also, we improved the alignment performances with the following steps:

(a) Filter the alignment read with mismatches ≤ 5 and mapping quality = 0.

(b) Correct the alignment using the package Picard (Supplemental URLs) with two core commands. Particularly, the ‘AddOrReplaceReadGroups’ command was used to replace all read groups in the INPUT file with a new read group and assign 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.

(c) Removed potential PCR duplication. If multiple read pairs had identical external coordinates, only the pair with the highest mapping quality was retained.

(d) Realigned the reads around indels by first identifying regions for realignment where at least one read contained an insertion or deletion with a

cluster of mismatching bases around it.

A total of 13.20 M candidate SNPs were used for these filtering steps and approximately ~ 0.71 M spurious SNPs were filtered and removed. Finally, ~12.48 M (94.55% of 13.20 M candidate SNPs) confident heterozygous or homozygous SNPs for each individual were located (depth ≥ 10) by reads mapping approach in assembled genomes and reference genome, simultaneously (**Fig. 1**).

5. SNP calling using the resequencing approach

After mapping reads against the reference genome (See **'4.2 Determination of the heterozygous or homozygous SNPs by aligning short sequencing reads'**), we performed SNP calling using two currently dominant algorithms (i.e. SAMtools[3] and GATK[4]).

5.1 SAMtools

The alignment was processed using a Bayesian approach as implemented in the package SAMtools (v.1.3)[3], which performed variant calling using the 'mpileup' program with the parameters '-C -D -S -m 2 -F 0.002 -d 1000'. Variants were strictly filtered for downstream analyses by requiring a minimum coverage of 10 and a maximum coverage of 1,000, a minimum RMS mapping quality of 20, the distance between adjacent SNPs ≥ 5 bp and no gaps present within a 3 bp window.

5.2 GATK tool

Base quality scores were recalibrated using the Genome Analysis Toolkit (GATK, v3.4)[4] with the analysis type of HaplotypeCaller based method of local *de novo* assembler and HMM likelihood function, which provides empirically accurate base quality scores for each base in every read. After SNP calling, we applied variant quality recalibration to exclude potential false positive variant calls. We used the command 'VariantFiltration' with the parameters '--filterExpression "QD < 10.0 || FS > 60.0 || MQ < 40.0 || ReadPosRankSum < -8.0" -G_filter "GQ<20".

Consequently, we detected 8.11 M and 7.77 M SNPs for each individual using SAMtools and GATK tool, respectively. Of which 7.41 M SNPs were

concurrently identified by the two algorithms, which accounted for 91.24% and 95.34% SNPs identified by SAMtools and GATK, respectively (**Fig. 1** and **Supplemental Fig. S8**). In addition, more than 98% and 97% of SNPs identified by SAMtools or GATK tool were also detected by assembly-versus-assembly method, respectively.

6. Extrapolation of the global SNP repertoire of all pig breeds

To obtain a reasonable estimate of the global SNP repertoire of pig species (i.e. the union and non-redundant SNPs presented in at least one of the pig individual/breed) (**Supplemental Fig. S14**), the number of newly added SNPs found on sequential addition of each new breed was extrapolated by fitting a two-phase exponential decay function to the data, where sequentially added breed number and the number of union SNPs of all the included breeds were considered as independent and dependent variables, separately.

We sequentially included up to ten breeds and simulated all the combinations of certain number of breeds: for n breeds included $N = 10! / [(10 - n)! \times n!]$ breed combinations were simulated, and the number of union SNPs in each breed combination was then calculated. When including only one breed, 10 data points were reported each presenting the total number of SNP in each breed. When including two breeds, $(10 \times 9) / 2 = 45$ data points were reported representing each possible pair that could be chosen from the ten breeds. When including three breeds, $(10 \times 9 \times 8) / (3 \times 2 \times 1) = 120$ data points were shown, corresponding to all possible three breed combinations that could be extracted from the ten breeds. By fitting the average values of union number of SNP as a function of the included breed number, the number of repertoire SNPs could be extrapolated. The average values of union number of SNP closely follow a two-phase exponential association that could be fitted. The number of union SNPs increased as more breeds were added. Nevertheless, extrapolation of the fitted curve demonstrated that with extremely high goodness-of-fit R square values 1.000.

7. Identification of insertions and deletions (indels)

7.1 Identification of candidate indels by assembly-versus-assembly method

We first identified candidate indels in ten assemblies using a previously reported methodology[156] (**Supplemental Table S9**). In brief, the *de novo* assembly of ten individuals were separately aligned onto the reference genome (*Sscrofa*10.2) using the LASTZ tool (Supplemental URLs) (See '**4.1 Gapped alignment of *de novo* assemblies to a reference genome'** for details**)**. The predicted gaps in the pair-wise alignments between the two genome assemblies were extracted using the SOAP software (Supplemental URLs) and defined as candidate indels.

7.2 Validation of confident indels by read mapping approach

To filter out spurious indels, we separately aligned the reads onto both the reference genome (*Sscrofa*10.2) and ten assemblies using the BWA tool[43], and calculated the read coverage for each candidate indel. Then the different criteria were used to validate the candidate indels ≤ 50 bp or > 50 bp as previously described[156]. The confident indels (≤ 50 bp) should be supported by more than three gapped aligned reads and their predicted breakpoints and/or genotype were perfectly consistent with the aligned reads. The confident indels (> 50 bp) should have significant differences in S/P ratios (i.e. the number of aligned single-end reads versus the number of aligned paired-end reads) between *de novo* assembly and reference genome ($P < 0.05$, Fisher's exact test), and be more than three times of the s.d. of the insert-size in length[156].

8. Copy number variation detection

We performed copy number variation (CNV) calling using the read depth method as implemented in the CNVnator tool[157]. To avoid noise from repeated elements in reference genome (Sscrofa 10.2) for the read depth method, we separately mapped the high quality short-insert reads (180 and 500 bp) of each individual (151.08 Gb) onto the repeat masked reference genome (59.95-fold coverage per individual) using the BWA software (v.0.7.12)[43]. To reduce the false positive discovery rate and avoid misinterpreting the results, we only retained CNVs longer than 1 kb and "N" content < 10% for subsequent analysis.

Consequently, we detected ~4,539 copy number gain events (a total of 23.00 Mb) for each breed (**Supplemental Table S15**). We only identified ~83 copy number loss events (a total of 0.42 Mb) for each breed. This dramatic reduction

of identified copy number loss events may be attributed to the increased “N” content in the considerable low quality (13.85%) and low coverage (26.6%) regions in reference genome[158], which were excluded by our criterion for a CNV (“N” content < 10%). Compared with copy number gain events, the identified copy number loss events exhibited significantly higher “N” content (7.21% versus 3.71%, $P < 10^{-16}$, Mann-Whitney *U* test). Errors in the reference assembly increase the number of false-positives and it was therefore decided to exclude copy number loss event from further analyses[159-161].

We further defined 15,914 copy number gain regions (a total of 72.27 Mb) by merging all overlapping calls across ten individuals into unique regions using a custom Perl script. Most putative copy number gain regions (72.74%, or 53.07 Mb) were found to be of low variability across ten breeds (coefficient of variation of copy number ≤ 10%). A total of 575 protein-coding genes were found to overlap (≥ 50%) with the 3,847 highly variable copy number gain regions among ten breeds (a total of 19.70 Mb, coefficient of variation of copy number > 10%), which mainly represented the highly variable gene family, such as olfactory receptors and immunoglobulin[159,161].

9. Functional enrichment analyses for genes

Functional enrichment analysis of Gene Ontology (GO) terms and pathways was performed using the DAVID[162] (Database for Annotation, Visualization and Integrated Discovery) web server (Supplemental URLs). Genes were mapped to their respective human orthologs, and the lists were submitted to DAVID for enrichment analysis of significant overrepresentation of GO biological processes (GO-BP), molecular function (GO-MF) terminologies, and categories of InterPro domain and KEGG-pathway. In all tests, the whole set of known genes was appointed as the background, and *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. Only GO-BP, GO-MF, KEGG-pathway or InterPro domain terms with a *P* value less than 0.05 were considered as significant and listed.

10. Identification of regions of homozygosity (ROHs)

The ROHs are defined as a continuous or uninterrupted stretch of a DNA

sequence without heterozygosity in the diploid state, regions of a minimum size of 200 kb encompassing 50 homozygous SNPs while allowing one heterozygous SNP were identified for each breed using PLINK (v.1.9)[163] (**Fig. 2B** and **Supplemental Fig. S12**).

11. Phylogenetic tree, PCA and LD analysis

The neighbor-joining phylogenetic tree was inferred using the package TreeBeST under the *p*-distances model (**Fig. 2B** and **Supplemental Figs. S24A,42**). PCA analysis was implemented in package EIGENSOFT (v.6.0.1)[164] (**Supplemental Figs. S13A** and **S24B**). To evaluate LD decay, the coefficient of determination (r^2) between any two loci was calculated using Haploview (v.4.2)[165] (**Fig. 4C** and **Supplemental Figs. S13,23,42**).

12. Calculation of identity score (IS)

Identity scores (ISs) were calculated to evaluate the pairwise similarities of the sequences (or assemblies) (**Figs. 2C,3B** and **Supplemental Figs. S25,30B**). For each identified SNP, we determined the fraction of reads that corresponded to the reference allele, *F*, in each sequence. The IS values of individual SNPs were then calculated as IS = 1 -($|F_{\text{sequence1}} - F_{\text{sequence2}}|$), with SNPs assessed only if at least one read was obtained in each sequence. The IS value for a sequence was the mean of all SNP IS values observed in the sequence for a specific comparison.

13. Identification of selected regions using RSD algorithm

To identify signatures of diversifying selection of pig breeds, relative homozygous SNP density (RSD) in nonoverlapping 10kb windows was calculated using a previously reported methodology[166] (**Fig. 3**, **Supplemental Figs. S23-25** and **Table S12**). The ungenotyped and/or heterozygous loci in at least one breed were filtered. Out of 259,511 non-overlapping 10 kb windows

across reference genome (Chromosomes 1 to 18 and X; a total of 211.87 Mb unplaced scaffolds in reference genome were excluded), 136,121 windows containing ≥ 10 homozygous SNPs were used to detect signatures of selective sweeps. Consequently, 4.60 M homozygous SNPs (out of a total of 5.00 M homozygous SNPs) were used for subsequent analyses.

The RSD value for each of 10 kb window was estimated for each breed utilizing the formula below:

$$\mathrm{RSD}_i = \frac{\text{Number of Homozygous SNP in breed } i}{\text{Total number of homozygous SNP}} \times 100$$

We randomly picked *k* breeds from the pool, where *k* can range from one to total breed number, and estimated observed and expected RSD values for each breed combination. Observed RSD for each of the *k* breeds can be calculated according to the above formula. The following are ideal cases exemplifying calculation of expected RSD values for each combination of breeds. If the combination only had one breed, we assumed ideally just this breed showed SNPs against the reference assembly, but all the other breeds showed the same homozygous alleles as the reference at each locus within the 10 kb window, so this breed would have RSD value 100 and the others would have zero. If the combination had two breeds, we would assume equal number of SNPs for each of them with all the others showing reference alleles, therefore these two breeds would each have RSD value 50 and the others would have zero. Similarly, for combination of three breeds, expected RSD values for each would be 33.33.

The χ^2 test value for each combination was then estimated using the following formula. It is worth noting that only the RSD values for the *k* breeds in each specific combination was involved in each calculation. The breed combination with the lowest χ^2 test value for each 10 kb window was considered as the best fit to the model of equal distribution of SNPs between breeds.

$$\mathrm{SC} = min\left(\binom{n}{k = 1 \ldots n} \sum_{i=1}^{k} \frac{(Obs_i - Exp_i)^2}{Exp_i} \right)$$

Successive 10 kb windows with the same breed combination were merged.

If one window had two successive windows of the same combination on each side, the five windows were merged and assigned the combination of four windows on both side. We estimated the probability that long genomic regions of the same combination appear after merging, by permuting (n = 10 million) the breed combination assigned to each window within each chromosome, and dividing the iterations when the maximum length of segment assigned to the same breed combination is longer than the observed length, by the total number of iterations. FDRtools (v.1.2.15)[167] was utilized to calculate FDR values corrected for multiple hypothesis test, and separate genomic regions with FDR less than 0.05 were considered as outliers.

14. Assigning missing sequences to the reference genome gaps

To incorporate the missing sequences to the reference genome assembly gaps (Sscrofa10.2), we retrieved 7.97 G 'One End Anchor (OEA) reads' for each individual from the paired-end long-insert (2, 5, 6 and 10 kb) libraries, where only one end can be uniquely mapped to the reference genome (**Supplemental Fig. S27**). These OEA reads were further mapped onto both their respective missing sequences (137.02 Mb per assembly) and the reference genome using the BWA software[43]. We counted the effectively anchored OEA reads, where one end of the pair was anchored to the missing sequences, and the other one was anchored to the flanking sequence at both ends of a large gap (maximal 1.5 folds of the library insert size, i.e., 3, 7.5, 9 and 15 kb in length for 2, 5, 6 and 10 kb insert size libraries, respectively). A missing sequence was assigned to a gap that had the most effectively anchored OEA reads (≥ 4) with this missing sequence. Out of 289.24 Mb gaps in reference pig assembly (the latest release, Sscrofa10.2), only 5,317 gaps (266.15 Mb in length) greater than or equal to 50 kb in length were used (**Supplemental Figs. S31**,**32** and **Table S16**).

15. RNA-seq

15.1 Samples for RNA-seq

Nine kinds of tissues (including *longissimus dorsi* muscle, *psoas* major muscle, subcutaneous adipose, heart, liver, spleen, lung, kidney and ovary) and a mixture of multiple tissues were obtained from the same individuals used for

each of ten *de novo* genome assemblies, exception of the lung tissue of Rongchang pig and the subcutaneous adipose, ovary, and mixture of multiple tissues of Tibetan wild boar. Total RNA was extracted using Trizol reagent (Life Technologies) according to the manufacturer's protocols. Sequencing libraries were generated following the manufacturer's standard procedures (Illumina). The 96 strand-specific RNA libraries were sequenced on the Illumina HiSeq 2500 platform, and generated a total of ~510.88 Gb high-quality data (100 bp paired-end reads). Bases were called using the Illumina CASAVA software. RNA-seq data of four tissues of Tibetan wild boar (heart, kidney, liver and lung) were downloaded from the NCBI-Gene Expression Omnibus (GEO) database under accession code GSE43892.

15.2. RNA-seq read mapping and data processing

High-quality reads were obtained by discarding reads that contained over 10% Ns (uncalled bases), had at least 10 bp that overlapped with adaptor sequences at mismatch rate equal to or less than 10%, or included over 50% of bases with base quality scores less than five. The remaining reads were aligned to the ten *de novo* assemblies (**Supplemental Figs. S35**,**46**) or reference genome (**Supplemental Fig. S38**) using TopHat (v.2.1.0) with default parameters[168]. Cufflinks (v.2.2.1) was used to quantify gene expression and obtain FPKM expression values (denoted as fragments per kb of transcript per Mb of mapped reads) with default parameters[169]. We carried out read alignment and expression quantification separately for each sample.

16. Structure annotation of protein coding genes in pig assemblies

To accurately explore the genes absent from the reference genome, we completely annotated the gene structures for ten assemblies by combining the evidences of the reference assembly-guided approach, the *ab initio*- and the homology-based methods, as well as transcription from the RNA-seq data (**Supplemental Fig. S33** and **Table S17**).

16.1 Reference assembly-guided annotation

To obtain high-quality gene models for each assembly, after repeat masking, we separately aligned ten assemblies to the reference assembly using LASTZ program (See '**4.1 Gapped alignment of *de novo* assemblies to a reference**

genome' for details**)**. The raw alignments were combined into larger blocks using the ChainNet algorithm; it generated a group of high-confidence pairwise syntenic blocks between the queried and the reference assembly. For the syntenic regions, the annotated genes in reference genome were directly mapped to each assembly and defined using GeneWise[170].

16.2 A*b initio*- prediction, the homology-based prediction, the transcription from the RNA-seq data

We also predicted structure of protein coding genes in ten assemblies using *ab initio*-, and homology-based methods, and by incorporating evidence of transcription from the RNA-seq data.

(a) *Ab initio* prediction

We used the *ab initio* predication packages Augustus[171], GeneID[172], Genscan[173], GlimmerHMM[174] and SNAP[175] with the parameters trained from a set of high-quality homologous prediction proteins.

(b) Homology-based prediction

The protein repertoires of human, mouse, sheep and cow were downloaded from Ensembl (Supplemental URLs) and mapped onto the repeat-masked assemblies using TBLASTn[176]. Then, homologous genome sequences were aligned against the matching proteins using Genewise[170] to define gene models.

(c) RNA-seq data

To enhance the genome annotation, 96 RNA libraries (7 to 10 libraries for each of ten individuals) were used. RNA-seq reads were aligned to their respective assembly using TopHat (v2.1.0)[168] with default parameters to identify exons region and splice positions. The alignment results were then used as input for Cufflinks (v2.2.1)[169] with default parameters for genome-based transcript assembly.

The non-redundant reference gene set was generated by merging genes predicted by three methods using EvidenceModeler (EVM, v.1.1.1)[177], and genes with ≤ 50 amino acids or only with *de novo* predictive support were removed. Of which, the gene models overlapped with those from the reference-

guided predication by aligning ten assemblies to the reference assembly were further filtered.

Consequently, we predicted an average of 20,782 protein-coding genes in each of the ten assemblies (**Supplemental Fig. S33** and **Table S17**).

17. Functional annotation of missing genes

We aligned the protein sequences of missing genes to NCBI RefSeqGene proteins of pig, human, cow and mouse (downloaded at June 13, 2015, Supplemental URLs). Gene functions were assigned according to the best match of the alignment to SwissProt and TEMBL databases[178], using BLASTp[176]. We annotated motifs and domains using InterPro[179] by searching against publicly available databases, including Pfam[180], PRINTS, PROSITE, ProDom, and SMART using InterProScan[179,181]. Gene Ontology (GO) terms[182] for each gene were retrieved from the corresponding InterPro descriptions. Furthermore, we mapped these missing genes to the KEGG pathway[183] to identify the best match category for each gene (**Supplemental Table S20**).

18. Calculating Shannon entropy as a measure of tissue specificity of missing genes

The Shannon entropy (H) was calculated with a custom Perl script using the formula described in original paper[184], which measures the degree of overall tissue specificity of a gene (**Supplemental Fig. S46**) The relative expression of a gene g in a tissue t relative to its expression given in N tissues is defined as:

$$P_{t|g} = W_{g,t} / \sum_{1 \le t \le N} W_{g,t}$$

where $w_{g,t}$ is the expression level of the gene g in tissue t. The Shannon entropy of a gene's expression distribution is then calculated as:

$$H_g = \sum_{1 \le t \le N} -P_{t|g} \log_2 P_{t|g}$$

This value is expressed in bits and ranges from zero to $\log_2(P_{t|g})$ genes expressed in a single tissue and uniformly expressed in all the common tissues examined, respectively.

19. Detecting coding SNPs in missing genes under selection

We performed F_{ST}-based approaches to investigate the selection signals in missing genes. We chose coding SNPs and excluded those with minor allele frequency less than 0.05. Chinese wild boars and each of seven Chinese domestic pig populations were pairwise-tested using the finite island model (with the *FDIST* approach) in Arlequin (v.3.5.2.2)[185] to detect outliers (**Supplemental Fig. S43**). Parameters were set as default values, with the exception of setting 200,000 simulations and allowing 10% missing data in each test. The *P* value for each locus was estimated using a kernel density approach. After completing the analysis, we performed multiple hypothesis test correction of *P* values utilizing FDRtools (v.1.2.15)[167]. We considered the loci with FDR less than 0.05 as possible outliers.