


RESEARCH ARTICLE

The de novo assembly of a European wild boar genome revealed unique patterns of chromosomal structural variations and segmental duplications

Jianhai Chen¹  | Jie Zhong¹ | Xuefei He¹ | Xiaoyu Li¹ | Pan Ni² | Toni Safner^{3,4} | Nikica Šprem³ | Jianlin Han^{5,6}

¹Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu, China

²Animal Husbandry and Veterinary Institute of Keqiao District, Shaoxing, Zhejiang, China

³Faculty of Agriculture, University of Zagreb, Zagreb, Croatia

⁴Centre of Excellence for Biodiversity and Molecular Plant Breeding, (CoE CroP-BioDiv), Zagreb, Croatia

⁵International Livestock Research Institute, Nairobi, Kenya

⁶CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China

Correspondence

Jianhai Chen, Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu 610041, China.
Email: jianhaichen@scu.edu.cn

Funding information

Fifth Batch of Technological Innovation Research Projects in Chengdu, Grant/Award Number: 2021-YF05-01331-SN; Postdoctoral Research and Development Fund of West China Hospital of Sichuan University, Grant/Award Number: 2020HXBH087

Abstract

The rapid progress of sequencing technology has greatly facilitated the de novo genome assembly of pig breeds. However, the assembly of the wild boar genome is still lacking, hampering our understanding of chromosomal and genomic evolution during domestication from wild boars into domestic pigs. Here, we sequenced and de novo assembled a European wild boar genome (ASM2165605v1) using the long-range information provided by 10× Linked-Reads sequencing. We achieved a high-quality assembly with contig N50 of 26.09 Mb. Additionally, 1.64% of the contigs (222) with lengths from 107.65 kb to 75.36 Mb covered 90.3% of the total genome size of ASM2165605v1 (~2.5 Gb). Mapping analysis revealed that the contigs can fill 24.73% (93/376) of the gaps present in the orthologous regions of the updated pig reference genome (Sscrofa11.1). We further improved the contigs into chromosome level with a reference-assistant scaffolding method. Using the ‘assembly-to-assembly’ approach, we identified intra-chromosomal large structural variations (SVs, length >1 kb) between ASM2165605v1 and Sscrofa11.1 assemblies. Interestingly, we found that the number of SV events on the X chromosome deviated significantly from the linear models fitting autosomes ($R^2 > 0.64$, $p < 0.001$). Specifically, deletions and insertions were deficient on the X chromosome by 66.14 and 58.41% respectively, whereas duplications and inversions were excessive on the X chromosome by 71.96 and 107.61% respectively. We further used the large segmental duplications (SDs, >1 kb) events as a proxy to understand the large-scale inter-chromosomal evolution, by resolving parental-derived relationships for SD pairs. We revealed a significant excess of SD movements from the X chromosome to autosomes ($p < 0.001$), consistent with the expectation of meiotic sex chromosome inactivation. Enrichment analyses indicated that the genes within derived SD copies on autosomes were significantly related to biological processes involving nervous system, lipid biosynthesis and

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Animal Genetics* published by John Wiley & Sons Ltd on behalf of Stichting International Foundation for Animal Genetics.

sperm motility ($p < 0.01$). Together, our analyses of the de novo assembly of ASM2165605v1 provides insight into the SVs between European wild boar and domestic pig, in addition to the ongoing process of meiotic sex chromosome inactivation in driving inter-chromosomal interaction between the sex chromosome and autosomes.

KEY WORDS

meiotic sex chromosome inactivation, reference genome, *Sus scrofa*, whole genome sequencing

INTRODUCTION

Pigs (*Sus scrofa domesticus*) were domesticated from their wild boar ancestors ~10,000 years ago during the early Neolithic agricultural revolution in the Near East and Central China independently (Rothschild & Ruvinsky, 2011). As one of the most important sources of animal protein, pork production has long been valued in East and Southeast Asia and some European countries. The pig industry is also expected to increase steadily at both global and regional scales in the next few years (<https://www.alliedmarketresearch.com/pork-meat-market>, accessed 20 December 2021). In addition to their valuable meat, pigs can also provide leather, bristles and lard products. The economically and agriculturally important features of pigs are majorly attributed to the long-lasting and ongoing efforts of breeding management and directional selection, especially since the Industrial Revolution (Bosse et al., 2014). Evidence based on comparative genomic analyses between wild boars and domestic pigs has shown the strong artificial selection that dramatically promoted the phenotypic transformation following domestication in both Europe and Asia (Li et al., 2010, 2017; Moon et al., 2015; Rubin et al., 2012; Wang et al., 2015; Yang et al., 2014).

In recent decades, pigs have also draw attention from medical field for translational research. Owing to their physiological and anatomical similarity to humans, pigs can serve as medical models for multiple human diseases. For example, there are reports on pig models for xenotransplantation (Blusch et al., 2002; Mariscal et al., 2018), wound healing (Sullivan et al., 2001), dental and orofacial research (Wang et al., 2007), gastrohelcoma (Tian et al., 2009), hearing loss (Guo & Yang, 2015) and neurodegenerative disorders (Holm et al., 2016). Research efforts in recent years have made tremendous strides toward the genome-wide editing of pigs, including the inactivation of porcine endogenous retroviruses (Yang et al., 2015) and germline engineering (Yue et al., 2021). These various medical explorations using pigs as large animal models strongly indicate the high potential of pigs in helping to tackle human medical problems.

Apart from their economic and medical value, pigs and their wild counterparts can also serve as a great model for evolution and population genetic studies. Just as J.B.S. Haldane, one of founders of population

genetics and neo-Darwin synthesis in 1930s and 1940s, proposed, 'One of the most hopeful fields for the study of evolution is the domestication of animals and perhaps also of plants' (Lickliter & Ness, 1990; Haldane, 1954). Even the establishment of Darwin's natural selection theory had been gleaned from extensive studies on the artificial selection of phenotypic variations in domesticated species (Darwin, 1875). In the current genomic era, domesticated species have been extensively sequenced, leading to the accumulation of abundant genomic data that are second to human population genomic data only. Unlike other popular domestic animals, whose wild counterparts are either limited in natural distribution (chickens and yak), endangered in population size (goats and sheep) or even extinct (horses and cattle), pigs are distinctive owing to their strong ability for widespread adaptation and long-range migration, thereby leading to their population flourishing for both wild boars and domestic breeds (Chen et al., 2018a; Johann et al., 2020; Rothschild & Ruvinsky, 2011).

The high-quality genome assemblies of domestic breeds/varieties have greatly promoted our understanding of numerous basic biological questions across a wide range of animals and plants, including the genetic bases of complex phenotypes in horses, chickens and pigs (Liu et al., 2022; Rubin et al., 2012; Wang et al., 2020), the ancient evolutionary processes of *S. scrofa* (Ai et al., 2015), genomic diversity in pigs (Li et al., 2017) and the origin and evolution of new genes in pigs and plants (Chen et al., 2019; Zhang et al., 2019b). Despite the fundamental role of genome assemblies in biological studies, high-quality assemblies are limited to well-known domestic pig breeds. For example, all currently available chromosome-level assemblies were from domestic breeds, including Duroc (Groenen et al., 2012; Warr et al., 2020), Bama (Zhang et al., 2019a), Luchuan (Yang et al., 2019), Ningxiang (Ma et al., 2022) and Meishan (Zhou et al., 2021), whereas the scaffold-level assemblies also came from domestic breeds, including Wuzhishan (Fang et al., 2012), Large White, Landrace, Berkshire, Hampshire, Pietrain, Bamei, Jinhua, Rongchang and Tibetan (Li et al., 2017). For wild boars, there are some short-read Illumina sequences (Bosse et al., 2015; Frantz et al., 2015; Groenen, 2016), but no assembly of a genome based on long-reads or long-range sequencing. In this study, we de novo assembled a European wild boar

genome using Linked-Reads sequencing (Marks et al., 2019; Weisenfeld et al., 2017; Zheng et al., 2016). This assembly provides insight into evolutionary patterns between wild boars and domestic pigs on both inter- and intra-chromosomal scales.

MATERIALS AND METHODS

DNA sampling, sequencing and assembly

Genomic DNA was extracted from the muscle tissue of a European male wild boar (France), which was collected during the regular hunting season according to national laws. The Linked-Reads approach developed by 10× Genomics was used for sequencing the genomic DNA (Marks et al., 2019). Briefly, the Linked-Reads method can provide long-range information for genomic short reads by leveraging microfluidics to partition and barcode the high-molecular-weight DNA. Following the recommendations of the sequencing platform (10× Genomics), we obtained ~56× depths of sequencing reads. The mitogenome was constructed using GetOrganelle (Jin et al., 2020) and used as a query to search the NCBI nucleotide database to confirm whether the sample was a wild boar. The de novo assembly of whole genome data was performed using SUPERNOVA v2.1.1 (Weisenfeld et al., 2017) downloaded from the official website of 10× Genomics, with default parameters. Finally, the contig-level assembly of ASM2165605v1 was upgraded into chromosomal scale with RAGTAG (Alonge et al., 2019).

Identification of structural variations

We identified structural variation (SVs), including deletions, insertions, duplications and inversions, using SYRI v1.4 (Goel et al., 2019). We further filtered out the short SVs of <1 kb and focused only on the continuous sequences for subsequent analyses. To understand whether the numbers of SVs were significant, we conducted regression analysis using the number of SVs against the length of relevant chromosomes. If the SVs are neutral and uniformly distributed, we may expect a linear pattern of the number of SVs conditional on chromosomal length. The regression analysis was performed with R packages. The biological processes of gene sets were analyzed using CLUSTERPROFILER (Wu et al., 2021).

Continuity analysis and mapping rate

To understand whether our new ASM2165605v1 assembly can fill the remaining gaps in the pig reference genome, Sscrofa11.1, we used MashMap aligner to achieve the 'one-to-one' syntenic region identification at first (identity over 90%) (Jain et al., 2018). BEDTOOLS software

(Quinlan & Hall, 2010) was further used to identify regions with gaps (represented by 'N') in Sscrofa11.1, but with uninterrupted sequences in ASM2165605v1. To understand whether our ASM2165605v1 can fill more gaps than previous assemblies of major European pig breeds, we repeated the above pipeline and compared the gap-filling rates among these assemblies. The assemblies of Asian pig breeds were not used to avoid potential misalignments. A comparison of mapping rates between different assemblies of European pig breeds and ASM2165605v1 was conducted using BWA-MEM (Li & Durbin, 2009). The marking of duplicate alignments was done using the SAMTOOLS suite (Li et al., 2009).

Genomic annotation for genes and repeats

We annotated ASM2165605v1 for its gene and repeat contents. The coding and non-coding genes were annotated by following the methods of previous assemblies (Groenen et al., 2012; Warr et al., 2020). Briefly, the protein-coding genes were annotated using the MAKER2 pipeline (Cantarel et al., 2008) by jointly using three methods, comprising RNAseq mapping, de novo predictions and homologous gene searching. The paired-end Illumina RNA-seq data of wild boars were downloaded from the BioProject of PRJEB3197 at NCBI. The de novo read mapping and assembly were conducted to obtain transcripts with packages of STAR (Dobin et al., 2013) and TRINITY (Haas et al., 2013). The genome-wide repeats were identified using REPEATMASKER with RM and Repbase repeats.

Identification of segmental duplications

We designed a pipeline, as visualized in Figure 4, to perform the identification of segmental duplications (SDs). Briefly, the whole genome alignment was performed by comparing Sscrofa11.1 against ASM2165605v1 using the LASTAL alignment tool (Hamada et al., 2017). The 'many-to-one' alignments over 1000 bp were kept as the domestic SD pair. To understand which one was the parental copy within a SD pair, we categorized the SDs into two types, the boundary-derived SD (bSD) and the internal-derived SD (iSD). For the iSD, it was easy to identify the parent-derived relationship, considering the feasible assumption that the synteny length of a parental copy should be longer than that of a derived copy. For the bSD, we determined the copying direction using BLASTN mapping of the SD pair against the orthologous copy in ASM2165605v1. The copy with a higher nucleotide identity of BLASTN comparison (>90%) was determined to be the parental copy because the distance between orthologous copies should be shorter than that between the derived copy and homologous copy. After determining the copying direction of SDs, we used the

linear regression to fit the number of SDs at the inter-chromosomal level. The CLUSTERPROFILER tool was used to conduct the over-representation test and gene-set enrichment analysis of Gene Ontology (Wu et al., 2021).

RESULTS

The sequencing and de novo assembling

The wild boar in this study was confirmed to be genetically nearest to European wild boar (FJ237002.1) based on complete mitogenomes with only two mismatches and a DNA identity of 99.9% (Figure S1). All other populations, including European local and commercial pigs as well as Asian wild boars and domestic pigs, showed a relatively low identity with the mitogenome assembled in this study.

For genomic data, in total, we obtained 1,696,695,959 linked reads, with 92.23% of them showing MapQ ≥ 30 . We generated the de novo assembly of the European wild boar, entitled ASM2165605v1, using SUPERNOVA (Weisenfeld et al., 2017), and kept the 13,542 contigs longer than 1000 bp. Among these contigs, there were 289 contigs longer than 100 kb, 77 contigs longer than 10 Mb and eight contigs longer than 50 Mb. The contig N50 value was 26.09 Mb, suggesting a high-level of continuity empowered by long-range information of the linked reads. Considering the close relationship between European wild boars and domestic pig breeds such as Large White, Berkshire, Landrace, Pietrain, Duroc and Hampshire (Frantz et al., 2015), we further examined whether ASM2165605v1 and contig-level assemblies of other European pig breeds can fill the gaps remaining in the current pig reference genome (Sscrofa11.1; Warr et al., 2020).

Rigorous 'one-to-one' orthologous mapping was conducted using MashMap (Jain et al., 2018) by focusing on orthologous segments with over 90% identity between the assemblies of European breeds (Li

et al., 2017), ASM2165605v1 and Sscrofa11.1. We revealed that the ASM2165605v1 contigs can fill more gaps in Sscrofa11.1 (93/376) than the current five assemblies of other European breeds (Figure 1a), suggesting that the continuity of the ASM2165605v1 assembly was better than those of all of the assemblies of other European breeds. Because extensive reports have established the close evolutionary relationship between local populations of wild boars and domestic breeds from Europe (Chen et al., 2018b; Frantz et al., 2013, 2015), the contigs of ASM2165605v1 could be ordered with the assistance of Sscrofa11.1 under the assumption of there being no large-scale inversion between the contigs. We anchored the contigs of ASM2165605v1 with the 'scaffold' function of RAGTAG (Alonge et al., 2019) and achieved a scaffold N50 of 4.24 kb after filtering out unplaced contigs. We further estimated sequence lengths for chromosomes with gapless DNA in ASM2165605v1 and compared them with the gapless lengths of Sscrofa11.1 using I_r ($I_r = \text{ASM2165605v1}_{\text{chr_length}} / \text{Sscrofa11.1}_{\text{chr_length}}$). We found I_r ratios ranging from 0.962 to 1.023 (Figure 1b), suggesting highly comparable genome coverages between ASM2165605v1 and Sscrofa11.1. Interestingly, chromosomes 5 and 10 were longer in ASM2165605v1 than in Sscrofa11.1, although all the remaining chromosomes demonstrated longer coverages in Sscrofa11.1 than in ASM2165605v1.

We also compared the mapping rates of six assemblies of European breeds (Sscrofa11.1, Hampshire, Pietrain, Landrace, Large White and Berkshire) and ASM2165605v1. We randomly chose the publicly available short-read DNA re-sequencing data of 10 European wild boars (Table S1) and mapped the cleaned reads to the six assemblies independently. We found that the mapping rates for all 10 wild boars were highest when using Sscrofa11.1 as a reference (median 97.64%), supporting the high quality of the most updated pig reference genome (Warr et al., 2020). In addition, ASM2165605v1 had a higher median mapping rate (97.45%) than the other five assemblies of domestic breeds except Sscrofa11.1,



FIGURE 1 (a) The number of gaps which are filled by the assemblies of ASM2165605v1 and other European pig breeds. (b) The comparison of non-missing lengths for all chromosomes of ASM2165605v1 and Sscrofa11.1 assemblies

suggesting its potential contribution to improving the overall continuity of pig pan-genomes.

To annotate the genome-wide protein-coding genes, we jointly applied three commonly used methods, including transcriptome alignment, de novo gene prediction and sequence homology-based predictions. In total, we obtained 21,400 protein-coding genes, which accounted for 1.3% of ASM2165605v1 (Table 1). We also annotated non-coding RNAs and genomic repeats (Table 2). In total, 0.273, 1.36, 0.131 and 0.77% of ASM2165605v1 was annotated as miRNA, tRNA, rRNA and snRNA respectively. Over 44% of ASM2165605v1 was identified as containing DNA repeats, including LINE, SINE, LTR, Satellite and unknown types of repeats, similar to previous reports on pig reference genome (Groenen et al., 2012; Warr et al., 2020).

The excess of inversions and duplications but the deficiency of deletions and insertions on the X chromosome

To understand the intra-chromosome variations based on structural variations (SVs), we compared ASM2165605v1 with Sscrofa11.1 using SyRI, which is a synteny and rearrangement identifier (Goel et al., 2019) (Figure 2 and Table S2). After removing SVs shorter than 1 kb and focusing only on the continuous sequences, we identified a total of 2700 SVs, including 1451 deletions, 833 insertions, 204 duplications and 212 inversions. Surprisingly, the longest inversion was found in chromosome 6 (1.49 Mb in Chr6:56947482-58549530 of Sscrofa11.1), harboring 52 protein-coding genes of six families inferred using the Markov Cluster Algorithm (van Dongen, 1991), of which only 12 had known functions (*ETFB*, *HAS1*, *LIM2*, *NKG7*, *PPP2R1A*, *SPACA6*, *VSIG10L*, *ZNF175*, *ZNF577*, *ZNF613*, *ZNF614* and *ZNF649*). Among these genes, *SPACA6* (sperm acrosome associated 6) was

reported to be required for fusion of sperm with the egg membrane during fertilization (Noda et al., 2020).

The highest number of SVs was present in chromosome 1, the longest one in the pig genome (Figure 3a). To understand whether the numbers of SVs on different chromosomes followed a uniform distribution model with the null hypothesis that the longer the chromosome is, the higher number of the SVs, we further analyzed the number of SVs (>1 kb) in a rigid statistical framework against the lengths of chromosomes. Interestingly, compared with all autosomes, the X chromosome was significantly deficient in deletions and insertions but excessive in duplications and inversions ($p < 0.001$; Figure 3b). This opposite pattern suggested that the X chromosome may have a different level of sensitivity for SVs affecting chromosomal structural or functional conservation. If we consider the differences in effective population sizes (N_e) between chromosomes, our observation is even more striking. The N_e of X chromosome was roughly three-quarters that of autosomes (Betrán et al., 2002), therefore the deviation of the X chromosome as an outlier would be even stronger. In detail, after adjusting the N_e estimates, the deficiency rates of deletions and insertions on the X chromosome were 66.14 and 58.41% respectively, whereas the excessive rates of duplications and inversions were relatively high, up to 71.96 and 107.61% respectively.

The excessive traffic of segmental duplications 'out of' instead of 'into' the X chromosome

As SVs only represent the intra-chromosomal variations, whether there are inter-chromosomal events involving large-scale segmental duplications (SDs) is still unknown. Here, we developed an in-house pipeline to identify the SDs between chromosomes (Figure 4). Based on the target chromosome, we defined two types

TABLE 1 The annotated protein-coding genes and non-coding genes in ASM2165605v1

Type	Subtype	Count	Average length (bp)	Total length (bp)	Percentage of genome
Coding genes		21,400	34,328	32,493,688	1.3014
miRNA		861	79	68,172	0.2730
tRNA		4471	76	338,344	1.3551
rRNA	rRNA	135	242	32,614	0.1306
	18S	9	1506	13,550	0.543
	28S	3	1610	4829	0.0193
	5.8S	6	154	925	0.0037
	5S	117	114	13,310	0.0533
snRNA	snRNA	1697	113	192,366	0.7705
	CD-box	294	92	26,999	0.1081
	HACA-box	277	135	37,347	0.1496
	Splicing	1100	112	123,581	0.004950
	scaRNA	26	171	4439	0.000178

TABLE 2 The annotated genomic repeats and their summaries in ASM2165605v1

Type	Rebase TEs		Other TEs		De novo		Combined TEs	
	Length (bp)	Percentage in genome	Length (bp)	Percentage in genome	Length (bp)	Percentage in genome	Length (bp)	Percentage in genome
DNA	74,618,563	2.99	3,922,533	0.16	28,749,938	1.15	76,142,807	3.05
LINE	486,903,714	19.5	227,213,753	9.1	551,923,837	22.11	665,121,705	26.64
SINE	22,342,313	0.89	0	0	25,909,379	1.04	35,962,746	1.44
LTR	132,835,795	5.32	6,568,546	0.26	2,616,79,411	10.48	318,961,338	12.77
Satellite	8,716,245	0.35	0	0	4,100,318	0.16	8,849,865	0.35
Unknown	1,147,190	0.05	9942	0	1,230,106	0.05	2,387,238	0.1
Total	726,563,820	29.1	237,714,774	9.52	873,592,989	34.99	1,107,425,699	44.35

Note: DNA refers to DNA transposons whereas LINE/SINE/LTR are retrotransposons. The TEs represents Transposable elements.

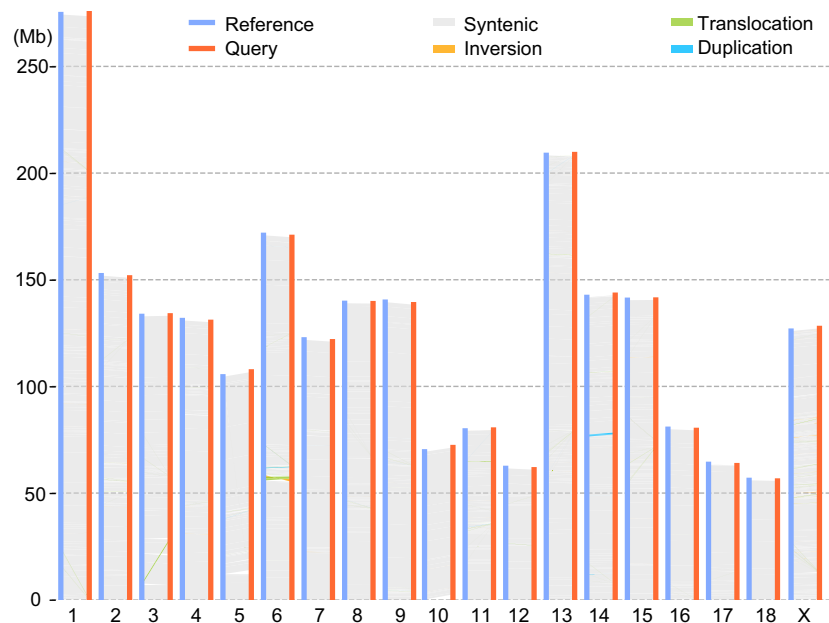


FIGURE 2 The structural variations between ASM2165605v1 and Sscrofa11.1 assemblies inferred with SyRI using default parameters. The four types of variations are shown in different colors

of copying directions between chromosomes, including 'into X' and 'into autosomes' (Figure 5). Additionally, based on the source/parental chromosomes, these inter-chromosomal types were further divided into three subtypes, which were 'autosomes to X (A>X)', 'autosomes to autosomes (A>A)', and 'X to autosomes (X>A)'. We found the 'A>X' subtype to be significantly shaped by a linear model for all autosomes ($R^2 = 0.75$, $p < 0.01$, blue in Figure 5). Likewise, the 'A>A' subtype also demonstrated a linear model ($R^2 = 0.81$, $p < 0.01$, red in Figure 5). In contrast, the 'X>A' subtype, an excessive outlier of the linear model, was significantly different from both 'A>A' and 'A>X' subtypes. These patterns suggested that X chromosome had served as an excessive source to 'export' SDs into autosomes.

Gene duplications can be roughly classified into two types, RNA- and DNA-mediated gene duplications, with the former arising through a mechanism termed retroposition or retroduplication (Kaessmann

et al., 2009), whereas the latter is processed by several mechanisms, including unequal cross-over and tandem, segmental, chromosomal and genome duplications (Kozlov, 2014). Previous reports have revealed the inter-chromosomal events of retrogenes (RNA-mediated gene duplications) in human, mouse, domestic pig and dog, and have found that X-derived retrogenes in autosomes are excessive (Betrán et al., 2002; Chen et al., 2019; Gao et al., 2019). There are also findings on the excess of X-derived genes on autosomes based on evidence of DNA-mediated gene duplication (Vibrantovski et al., 2009). The best-known hypotheses to explain this underlying preference of the X-derived movement involve sexual antagonism (Wu & Xu, 2003; Wyman et al., 2012) and meiotic sex chromosome inactivation (MSCI; Dai et al., 2006; Turner, 2007). Meiotic sex chromosome inactivation has been supported by a mouse model experiment, in which the evolutionarily new gene on an autosome can compensate for the

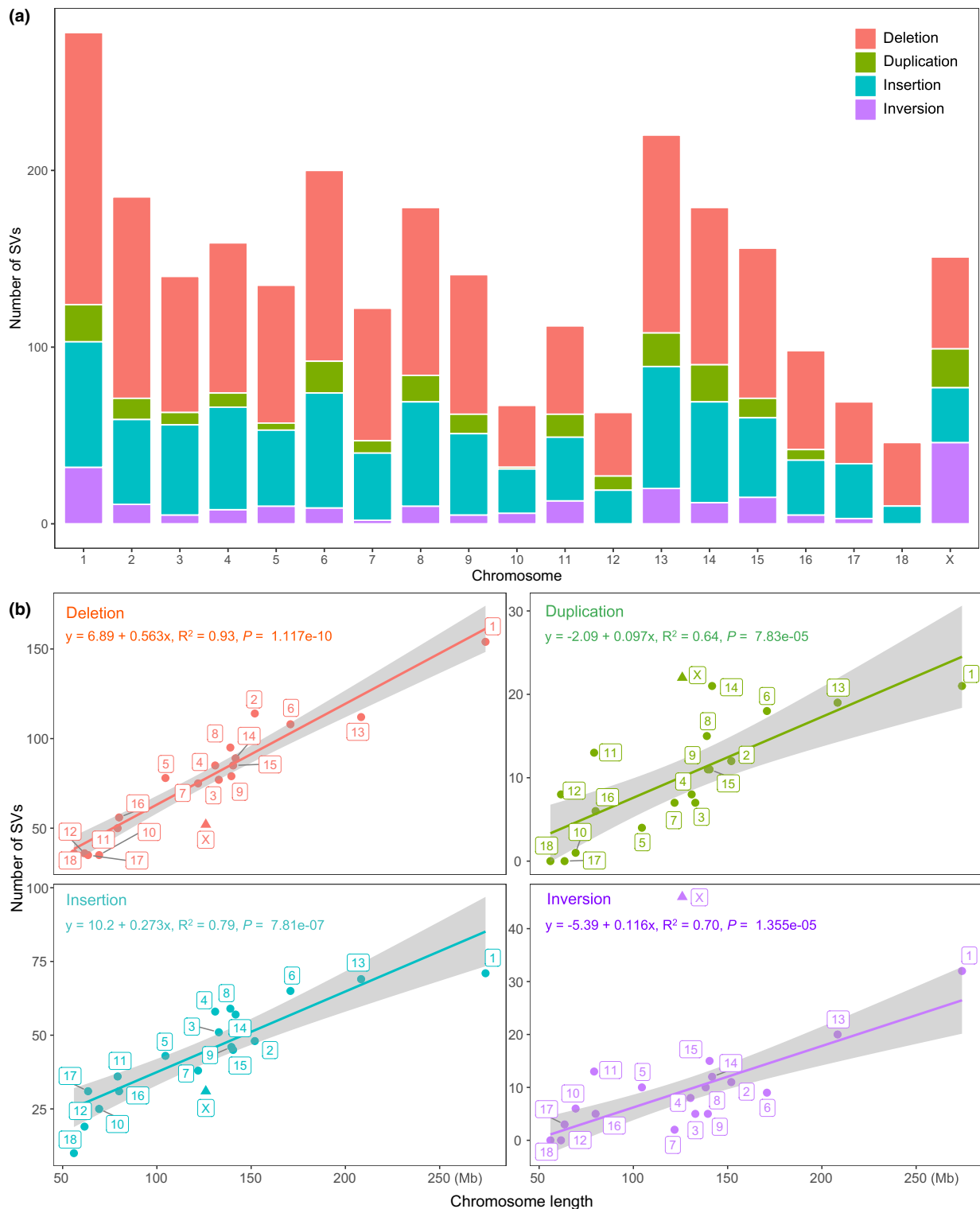


FIGURE 3 The numbers of structural variations across chromosomes (a) and the regression of the numbers of structural variations against the lengths of chromosomes (b)

function of parental gene in X chromosome owing to epigenetic silencing during the male meiosis (Jiang et al., 2017). In this study, we provided clues that the excess of X-derived SDs in autosomes could also be attributed to these molecular mechanisms, including MSC1.

If the hypothesis of MSC1 driving the excess of X-derived SDs is solid, we may expect that the genes covered by X-derived SDs are involved in male meiosis-related processes. Our enrichment analysis found that the genes linked with X-derived SDs were significantly

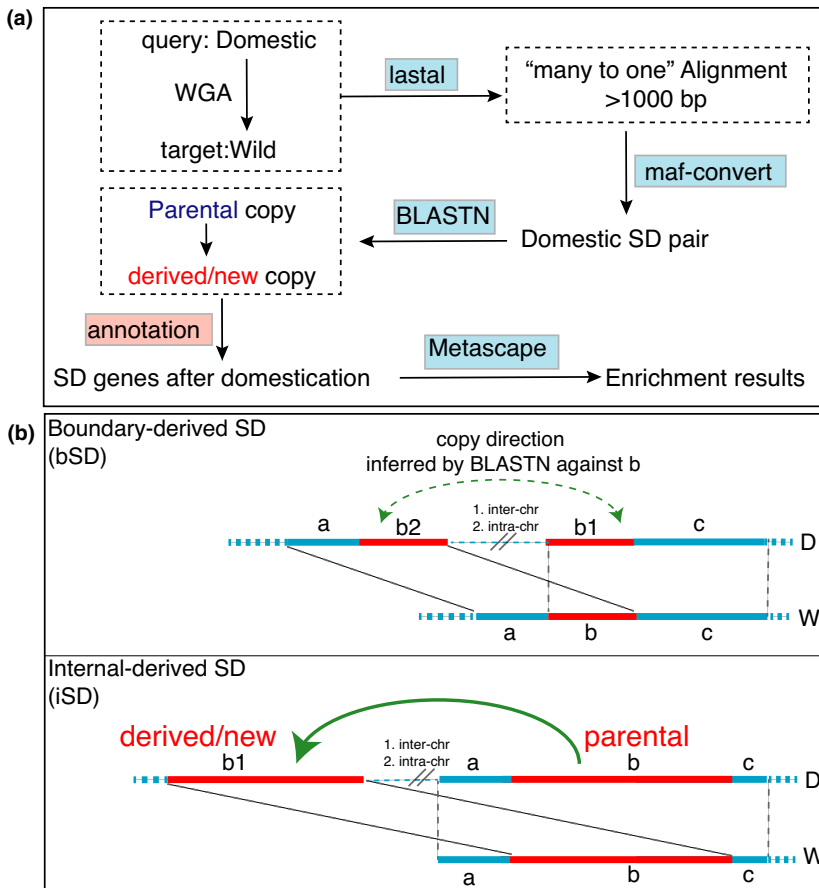


FIGURE 4 The pipeline designed for identifying the segmental duplications (SDs). (a) The flowchart of major software used and the overall processes. (b) The two types of SDs, which cover the boundary-derived SD (bSD) and internal-derived SD (iSD)

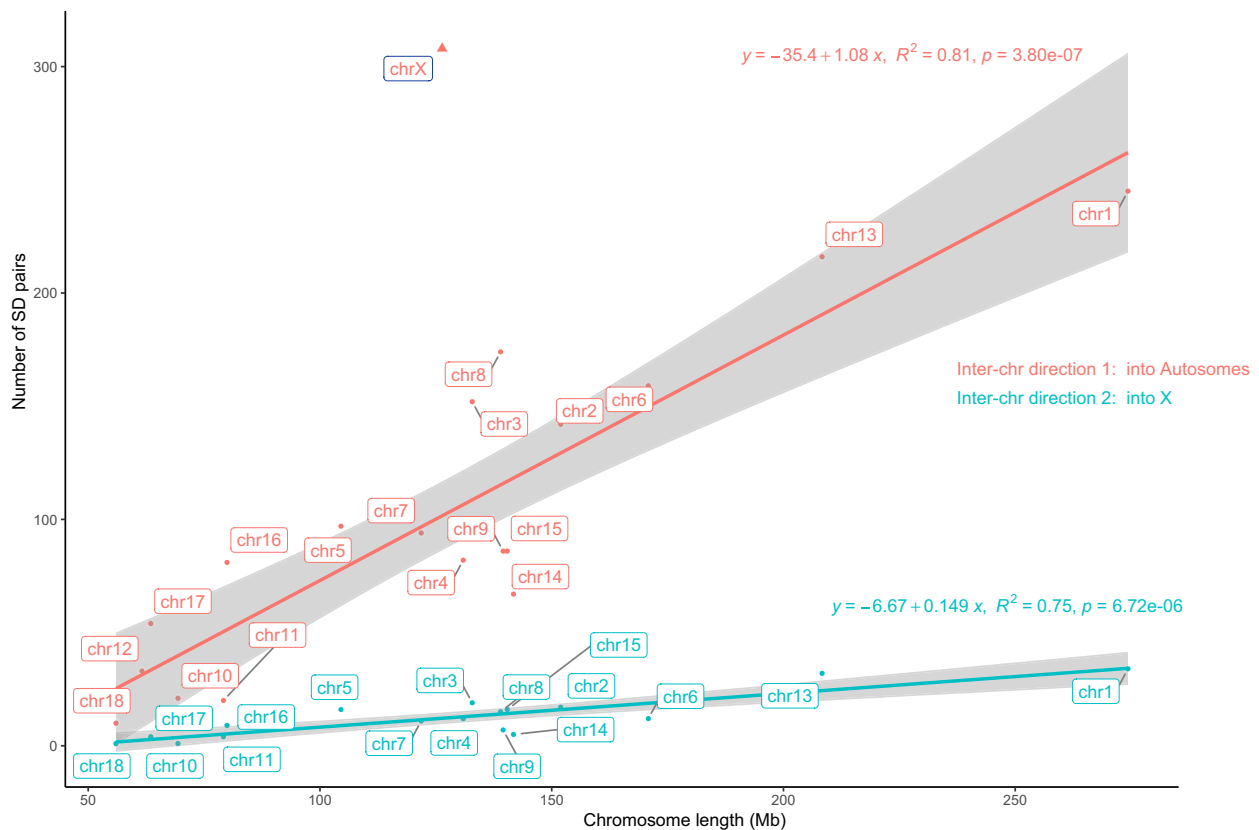


FIGURE 5 The regression of the numbers of SDs against the lengths of chromosomes. All numbers are inter-chromosomal SD numbers. Red and blue show the directions 'into autosomes' and 'into X chromosome', respectively

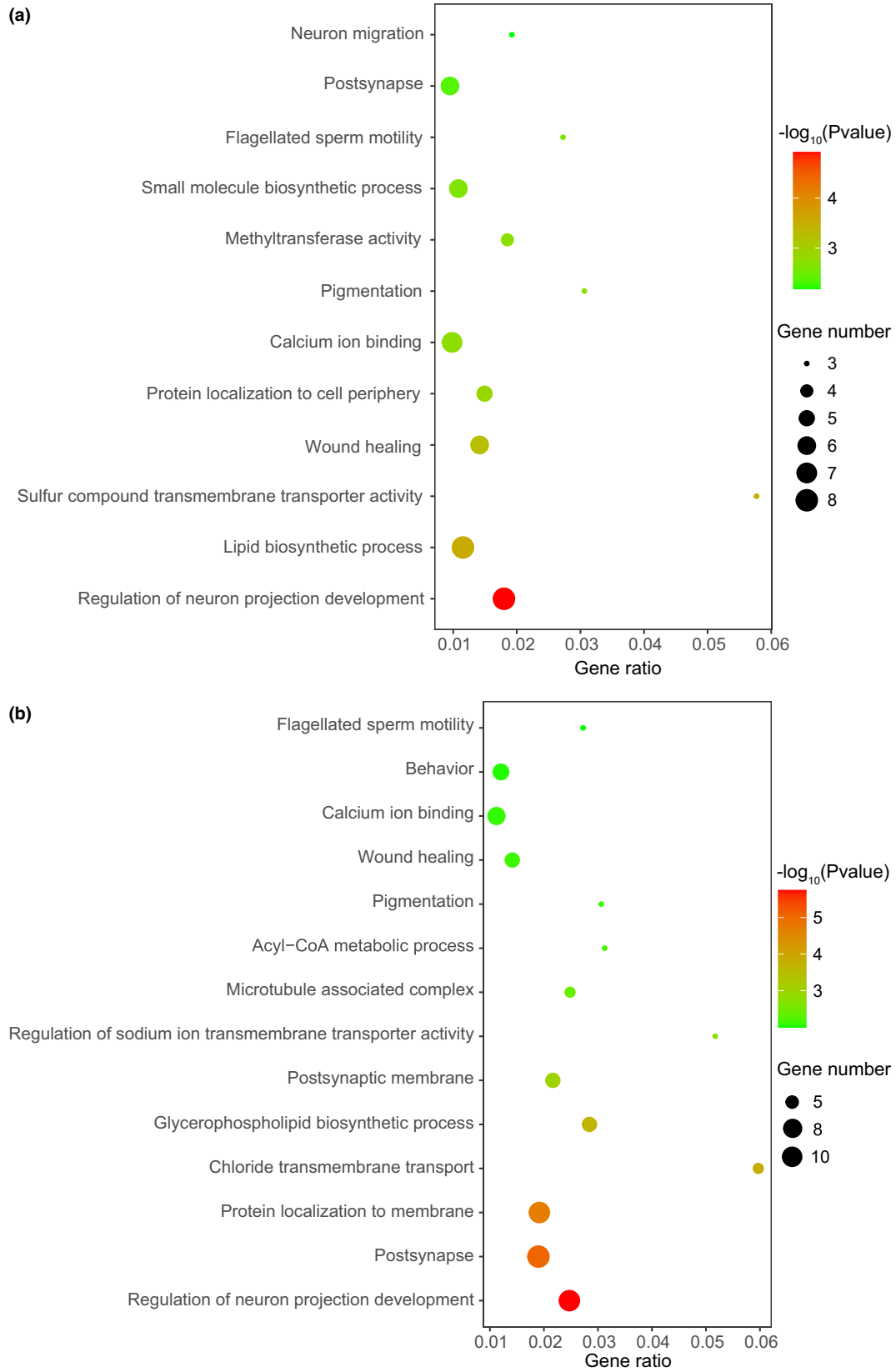


FIGURE 6 The enrichment analysis of biological processes using X-derived autosomal genes (a) and all genes (b) related to SD movements. All processes are statistically significant ($p < 0.01$) as visualized with colors from green to red

($p < 0.01$) enriched in multiple processes involving the nervous system, metabolism and reproductive system (Figure 6). The enriched processes were stable for both X-derived autosomal genes (Figure 6a) and all parental X-genes and derived-autosomal genes (Figure 6b). Specifically, the enriched biological process involving the reproductive system is flagellated sperm motility. This observation is probably relevant to MSC1, in which the epigenetic silence during mid- and post-meiosis may impose an evolutionary force to drive the male-meiotic advantageous genes to be transposed and expressed in autosomes.

DISCUSSION

Exploring the advantage of the 10× Linked-Reads sequencing, we de novo assembled the first, high-quality genome, ASM2165605v1, of a European wild boar with contig N50 of 26.09 Mb. The contents of genes, repeats and non-coding RNAs were highly similar between ASM2165605v1 and Sscrofa11.1. Notably, we recognized that, compared with the assemblies of several major European pig breeds stored in the Ensembl database, our ASM2165605v1 can fill the highest number of gaps in Sscrofa11.1. Overall, this novel ASM2165605v1 can therefore provide additional variations for the burgeoning pan-genomes of wild boars and domestic pigs.

Comparative analyses between ASM2165605v1 and Sscrofa11.1 revealed an interesting pattern of SVs. Statistically, the deletions and insertions were deficient, whereas the duplications and inversions were excessive on the X chromosome. This finding is insightful for us to understand the intra-chromosomal evolution at species level. Under the framework of the neutral evolution theory, we may expect the near linear distribution of SVs in chromosomes dependent on their lengths. Here, our observation of significant deficiency of the deletions and insertions in X chromosome suggests that this type of SV is under a stronger purifying selection than duplications and inversions. In contrast, as diversity is the genetic basis for positive selection, the excess duplications and inversions on the X chromosome advocate that they may have more chances to serve as a source of genetic variations for natural or artificial selection. Thus, our results support the selective heterogeneity of SVs on the X chromosome.

Meiotic sex chromosome inactivation is predicted to be an evolutionarily ancient mechanism critical for male reproductive processes. Owing to the importance of reproduction performance in domestic pigs, the domestication process provides a unique opportunity to test the impact of MSC1 in this species. Here, we identified the frequent SDs by comparing ASM2165605v1 and Sscrofa11.1 assemblies and revealed a significant excess of SDs copied from the X chromosome to autosomes. Previous reports have proposed and validated the process of MSC1, which can drive the relocation of genes

from the X chromosome to autosomes, to avoid the male meiotic silence of the X chromosome at both species level (Emerson et al., 2004; Jiang et al., 2017) and population level (Zhang & Tautz, 2021). Our observation is consistent with this well-accepted theory.

In summary, we generated, for the first time, the de novo assembly of a European wild boar, to provide a basic genomic resource for future studies, and to improve, deepen and widen our understanding on genome evolution during domestication. Regardless of the questions on genomic diversity, population variations or even multiple evolutionary processes, the novel set of SVs and SDs identified from the comparison of two high-quality wild boar and domestic pig assemblies may serve as an entry point for further exploration.

ACKNOWLEDGEMENTS

This study was supported by the Fifth Batch of Technological Innovation Research Projects in Chengdu (2021-YF05-01331-SN) and Postdoctoral Research and Development Fund of West China Hospital of Sichuan University (2020HXBH087).

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

The sequence data and genome assembly of ASM2165605v1 can be accessed through NCBI GenBank BioProject code [PRJNA791558](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA791558) and assembly accession no. GCA_021656055.1 respectively.

ORCID

Jianhai Chen  <https://orcid.org/0000-0002-7278-8090>

REFERENCES

- Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L. et al. (2015) Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics*, 47, 217–225.
- Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F.J. et al. (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, 20, 224.
- Betrán, E., Thornton, K. & Long, M. (2002) Retroposed new genes out of the X in *Drosophila*. *Genome Research*, 12, 1854–1859.
- Blusch, J.H., Patience, C. & Martin, U. (2002) Pig endogenous retroviruses and xenotransplantation. *Xenotransplantation*, 9, 242–251.
- Bosse, M., Megens, H.-J., Frantz, L.A.F., Madsen, O., Larson, G., Paudel, Y. et al. (2014) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature Communications*, 5, 1–8.
- Bosse, M., Megens, H.-J., Madsen, O., Crooijmans, R.P., Ryder, O.A., Austerlitz, F. et al. (2015) Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. *Genome Research*, 25, 970–981.
- Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B. et al. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18, 188–196.

- Chen, J., Mortola, E., Du, X., Zhao, S. & Liu, X. (2019) Excess of retrogene traffic in pig X chromosome. *Genetica*, 147, 23–32.
- Chen, J., Ni, P., Tran Thi, T.N., Kamalidinov, E.V., Petukhov, V.L., Han, J. et al. (2018a) Selective constraints in cold-region wild boars may defuse the effects of small effective population size on molecular evolution of mitogenomes. *Ecology and Evolution*, 8, 8102–8114.
- Chen, J., Ni, P., Li, X., Han, J., Jakovčić, I., Zhang, C. et al. (2018b) Population size may shape the accumulation of functional mutations following domestication. *BMC Evolutionary Biology*, 18, 4.
- Dai, H., Yoshimatsu, T.F. & Long, M. (2006) Retrogene movement within-and between-chromosomes in the evolution of *Drosophila* genomes. *Gene*, 385, 96–102.
- Darwin, C. (1875) *The variation of animals and plants under domestication*, 2nd edition. London: John Murray, Albemarle Street.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S. et al. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15–21.
- van Dongen, S. (1991) A cluster algorithm for graphs. Mathematics Subject Classification: 05B20, 05B25, 60J15, 62H30, 68T10, 90C35. 41 pp.
- Emerson, J., Kaessmann, H., Betrán, E. & Long, M. (2004) Extensive gene traffic on the mammalian X chromosome. *Science*, 303, 537–540.
- Fang, X., Mou, Y., Huang, Z., Li, Y., Han, L., Zhang, Y. et al. (2012) The sequence and analysis of a Chinese pig genome. *GigaScience*, 1, 16.
- Frantz, L.A.F., Schraiber, J.G., Madsen, O., Megens, H.-J., Bosse, M., Paudel, Y. et al. (2013) Genome sequencing reveals fine scale diversification and reticulation history during speciation in *Sus*. *Genome Biology*, 14, R107.
- Frantz, L.A., Schraiber, J.G., Madsen, O., Megens, H.-J., Cagan, A., Bosse, M. et al. (2015) Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nature Genetics*, 47, 1141–1148.
- Gao, X., Li, Y., Adetula, A.A., Wu, Y. & Chen, H. (2019) Analysis of new retrogenes provides insight into dog adaptive evolution. *Ecology and Evolution*, 9, 11185–11197.
- Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, 20, 277.
- Groenen, M.A.M. (2016) A decade of pig genome sequencing: a window on pig domestication and evolution. *Genetics Selection Evolution*, 48, 23.
- Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F. et al. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, 491, 393–398.
- Guo, W. & Yang, S.-M. (2015) Advantages of a miniature pig model in research on human hereditary hearing loss. *Journal of Otology*, 10, 105–107.
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J. et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8, 1494–1512.
- Haldane, J.B.S. (1954) *The statics of evolution*. Evolution as a process. London: George Allen and Unwin, 109–121.
- Hamada, M., Ono, Y., Asai, K. & Frith, M.C. (2017) Training alignment parameters for arbitrary sequencers with LAST-TRAIN. *Bioinformatics*, 33, 926–928.
- Holm, I.E., Alstrup, A.K.O. & Luo, Y. (2016) Genetically modified pig models for neurodegenerative disorders. *Journal of Pathology*, 238, 267–287.
- Jain, C., Koren, S., Dilthey, A., Phillippy, A.M. & Aluru, S. (2018) A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics*, 34, i748–i756.
- Jiang, L., Li, T., Zhang, X., Zhang, B., Yu, C., Li, Y. et al. (2017) RPL10L is required for male meiotic division by compensating for RPL10 during meiotic sex chromosome inactivation in mice. *Current Biology*, 27, 1498–1505.e6.
- Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C.W., Yi, T.-S. et al. (2020) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, 21, 241.
- Johann, F., Handschuh, M., Linderoth, P., Dormann, C.F. & Arnold, J. (2020) Adaptation of wild boar (*Sus scrofa*) activity in a human-dominated landscape. *BMC Ecology*, 20, 4.
- Kaessmann, H., Vinckenbosch, N. & Long, M. (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics*, 10, 19–31.
- Kozlov, A.P. (2014) Chapter 8 – The general principles and molecular mechanisms of the origin of novel genes. In: Kozlov, A.P. (Ed.) *Evolution by tumor neofunctionalization*. San Diego, CA: Academic Press, pp. 75–86.
- Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
- Li, J., Yang, H., Li, J., Li, H., Ning, T., Pan, X. et al. (2010) Artificial selection of the melanocortin receptor 1 gene in Chinese domestic pigs during domestication. *Heredity*, 105, 274–281.
- Li, M., Chen, L., Tian, S., Lin, Y.U., Tang, Q., Zhou, X. et al. (2017) Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Research*, 27, 865–874.
- Lickliter, R. & Ness, J.W. (1990) Domestication and comparative psychology: status and strategy. *Journal of Comparative Psychology*, 104, 211.
- Liu, X., Zhang, Y., Liu, W., Li, Y., Pan, J., Pu, Y. et al. (2022) A single-nucleotide mutation within the TBX3 enhancer increased body size in Chinese horses. *Current Biology*, 32, 480–487.e6.
- Ma, H., Jiang, J., He, J., Liu, H., Han, L., Gong, Y. et al. (2022) Long-read assembly of the Chinese indigenous Ningxiang pig genome and identification of genetic variations in fat metabolism among different breeds. *Molecular Ecology Resources*, 1–13.
- Mariscal, A., Caldarone, L., Tikkanen, J., Nakajima, D., Chen, M., Yeung, J. et al. (2018) Pig lung transplant survival model. *Nature Protocols*, 13, 1814–1828.
- Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R. et al. (2019) Resolving the full spectrum of human genome variation using linked-reads. *Genome Research*, 29, 635–645.
- Moon, S., Kim, T.-H., Lee, K.-T., Kwak, W., Lee, T., Lee, S.-W. et al. (2015) A genome-wide scan for signatures of directional selection in domesticated pigs. *BMC Genomics*, 16, 130.
- Noda, T., Lu, Y., Fujihara, Y., Oura, S., Koyano, T., Kobayashi, S. et al. (2020) Sperm proteins SOF1, TMEM95, and SPACA6 are required for sperm–oocyte fusion in mice. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 11493–11502.
- Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842.
- Rothschild, M.F. & Ruvinsky, A. (2011) *The genetics of the pig*, 2nd edition. Wallingford, Oxfordshire, UK: CABI.
- Rubin, C.-J., Megens, H.-J., Barrio, A.M., Maqbool, K., Sayyab, S., Schwochow, D. et al. (2012) Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 19529–19536.
- Sullivan, T.P., Eaglstein, W.H., Davis, S.C. & Mertz, P. (2001) The pig as a model for human wound healing. *Wound Repair and Regeneration*, 9, 66–76.
- Tian, Y.-Y., Luo, M.-F., Xu, Y.-H., Zhang, W.-B., Wang, G.-J., Wang, R.-H. et al. (2009) Mini-pig model of Gastrohelcoma induced

- by different irritant reagents. *Laboratory Animal Science*, 26(4), 32–34.
- Turner, J.M.A. (2007) Meiotic sex chromosome inactivation. *Development*, 134, 1823–1831.
- Vibrantovski, M.D., Zhang, Y. & Long, M. (2009) General gene movement off the X chromosome in the *Drosophila* genus. *Genome Research*, 19, 897–903.
- Wang, C., Wang, H., Zhang, Y., Tang, Z., Li, K. & Liu, B. (2015) Genome-wide analysis reveals artificial selection on coat colour and reproductive traits in Chinese domestic pigs. *Molecular Ecology Resources*, 15, 414–424.
- Wang, M.-S., Thakur, M., Peng, M.-S., Jiang, Y.U., Frantz, L.A.F., Li, M. et al. (2020) 863 genomes reveal the origin and domestication of chicken. *Cell Research*, 30, 693–701.
- Wang, S., Liu, Y., Fang, D. & Shi, S. (2007) The miniature pig: a useful large animal model for dental and orofacial research. *Oral Diseases*, 13, 530–537.
- Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D.M., Billis, K. et al. (2020) An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience*, 9, giaa051.
- Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M. & Jaffe, D.B. (2017) Direct determination of diploid genome sequences. *Genome Research*, 27, 757–767.
- Wu, C.-I. & Xu, E.Y. (2003) Sexual antagonism and X inactivation – the SAXI hypothesis. *Trends in Genetics*, 19, 243–247.
- Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z. et al. (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation*, 2, 100141.
- Wyman, M.J., Cutter, A.D. & Rowe, L. (2012) Gene duplication in the evolution of sexual dimorphism. *Evolution*, 66, 1556–1566.
- Yang, L., Guell, M., Niu, D., George, H., Lesha, E., Grishin, D. et al. (2015) Genome-wide inactivation of porcine endogenous retroviruses (PERVs). *Science*, 350, 1101–1104.
- Yang, S., Li, X., Li, K., Fan, B. & Tang, Z. (2014) A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. *BMC Genetics*, 15, 7.
- Yang, Y., Lian, J., Xie, B., Chen, M., Niu, Y., Li, Q. et al. (2019) *Chromosome-scale de novo assembly and phasing of a Chinese indigenous pig genome*. <https://doi.org/10.1101/770958>
- Yue, Y., Xu, W., Kan, Y., Zhao, H.-Y., Zhou, Y., Song, X. et al. (2021) Extensive germline genome engineering in pigs. *Nature Biomedical Engineering*, 5, 134–143.
- Zhang, L., Huang, Y., Wang, M., Guo, Y., Liang, J., Yang, X. et al. (2019a) Development and genome sequencing of a laboratory-inbred miniature pig facilitates study of human diabetic disease. *iScience*, 19, 162–176.
- Zhang, L.I., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A.R. et al. (2019b) Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nature Ecology & Evolution*, 3, 679–690.
- Zhang, W. & Tautz, D. (2021) Tracing the origin and evolutionary fate of recent gene retrocopies in natural populations of the house mouse. *Molecular Biology and Evolution*, 39, msab360.
- Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M. et al. (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34, 303–311.
- Zhou, R., Li, S.-T., Yao, W.-Y., Xie, C.-D., Chen, Z., Zeng, Z.-J. et al. (2021) The Meishan pig genome reveals structural variation-mediated gene expression and phenotypic divergence underlying Asian pig domestication. *Molecular Ecology Resources*, 21, 2077–2092.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

Fig S1

Tables S1 and S2

How to cite this article: Chen, J., Zhong, J., He, X., Li, X., Ni, P., Safner, T., et al (2022) The de novo assembly of a European wild boar genome revealed unique patterns of chromosomal structural variations and segmental duplications. *Animal Genetics*, 53, 281–292. Available from: <https://doi.org/10.1111/age.13181>