Chromosome-scale genomes provide new insights into subspecies divergence and evolutionary characteristics of the giant panda

Guang, Xuanmin; Lan, Tianming; Wan, Qiu Hong; Huang, Yan; Li, Hong; Zhang, Mingchun; Li, Rengui; Zhang, Zhizhong; Lei, Yinghu; Zhang, Ling; Zhang, Heming; Li, Desheng; Li, Xiaoping; Li, Haimeng; Xu, Yan; Qiao, Majiu; Wu, Daifu; Tang, Keyi; Zhao, Pengpeng; Lin, Jian-Qing; Sahu, Sunil Kumar; Liang, Qiqi; Jiang, Wenkai; Zhang, Danhui; Xu, Xun; Liu, Xin; Lisby, Michael; Yang, Huanming; Kristiansen, Karsten; Liu, Huan; Fang, Sheng-Guo

Published in: Science Bulletin

DOI: 10.1016/j.scib.2021.02.002

Publication date: 2021

Document version: Publisher's PDF, also known as Version of record

Document license: CC BY-NC-ND

Chromosome-scale genomes provide new insights into subspecies divergence and evolutionary characteristics of the giant panda

Xuanmin Guanga,b,1, Tianming Lanb,g,1, Qiu-Hong Wan a,1, Yan Huang c,1, Hong Li d,1, Mingchun Zhang c, Renguil,i, Zhizhong Zhang c, Yinhui Lei e, Ling Zhang f, Heming Zhang c, Desheng Li c, Xiaoping Li b,g, Haimeng Li b, Yan Xu b, Maiju Qiao c, Daifu Wu c, Keyi Tang a, Pengpeng Zhao e, Jian-Qing Lin a, Sunil Kumar Sahu b, Qiqi Liang d, Wenkai Jiang d, Danhui Zang e, Xin Xub,h, Xin Liub, Michael Lisbyg, Huanming Yangb,i, Sheng-Guo Fang a,*

*MOE Key Laboratory of BioSystems Homeostasis & Protection, State Conservation Centre for Gene Resources of Endangered Wildlife, College of Life Sciences, Zhejiang University, Hangzhou 310058, China
bState Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
cKey Laboratory of State Forestry and Grassland Administration (State Park Administration) on Conservation Biology of Rare Animals in the Giant Panda National Park, China Conservation and Research Center for the Giant Panda, Dujiaoyang 611830, China
dNovogene Bioinformatics Institute, Beijing 100083, China
eQingdao-Europe Advanced Institute for Life Sciences, Qingdao 266555, China
fGuangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen 518120, China
gGuangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen 518120, China
hLaboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark
iLaboratory of Genetics and Molecular Biomedicine, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark
jChina Wildlife Conservation Association, Beijing 100714, China
kQinling Research Center of Giant Panda Breeding, Shaanxi Academy of Forestry, Xi’an 710082, China
lUniversity, Hangzhou 310058, China
mState Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
nChina Wildlife Conservation Association, Beijing 100714, China
oKey Laboratory of State Forestry and Grassland Administration (State Park Administration) on Conservation Biology of Rare Animals in the Giant Panda National Park, China Conservation and Research Center for the Giant Panda, Dujiaoyang 611830, China
pMOE Key Laboratory of BioSystems Homeostasis & Protection, State Conservation Centre for Gene Resources of Endangered Wildlife, College of Life Sciences, Zhejiang University, Hangzhou 310058, China
qState Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
rKey Laboratory of State Forestry and Grassland Administration (State Park Administration) on Conservation Biology of Rare Animals in the Giant Panda National Park, China Conservation and Research Center for the Giant Panda, Dujiaoyang 611830, China
sNovogene Bioinformatics Institute, Beijing 100083, China
tQingdao-Europe Advanced Institute for Life Sciences, Qingdao 266555, China
uGuangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen 518120, China
vGuangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen 518120, China
wLaboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark
xLaboratory of Genetics and Molecular Biomedicine, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark
yChina Wildlife Conservation Association, Beijing 100714, China
zQinling Research Center of Giant Panda Breeding, Shaanxi Academy of Forestry, Xi’an 710082, China
{Corresponding authors.
E-mail addresses: kk@bio.ku.dk (K. Kristiansen), liuhuan@genomics.cn (H. Liu), sgfanglab@zju.edu.cn (S.-G. Fang).
1 These authors contributed equally to this work.

1. Introduction

The giant panda (Ailuropoda melanoleuca) is an iconic endangered species that has attracted considerable interest from the public and scientists. Extensive breeding programs have success-
The giant panda is traditionally divided into the Sichuan (Ailuropoda melanoleuca melanoleuca, SC) and the Qinling (Ailuropoda melanoleuca qinlingensis, QLI) subspecies based on pelage coloration and the statistical analysis of cranial and dental morphology [4,5] (Fig. 1a, b). However, the effective population sizes of the two subspecies appear to be extremely unbalanced. The QLI subspecies, which lives in Shaanxi Province, is considered to be the ancestral population of extant giant pandas [5]. Its population size is five times smaller than that of the SC subspecies. The adaptive and reproductive capacities of the SC subspecies are also superior to those of the QLI subspecies [5,6]. Determining how these differences developed and subsequently were maintained is important to improve conservation strategies for both giant panda subspecies.

Genome sequence and structure provide valuable information about the genetic mechanisms underlying distinct phenotypes and evolutionary trajectories. The previously reported draft genome assemblies of the giant panda genome from 2010 and 2019 indicated a close relationship between the bamboo diet and the omé assemblies of the giant panda genome from 2010 and 2019 and evolutionary trajectories. The previously reported draft genome sequence and structure provide valuable information about the genetic mechanisms underlying distinct phenotypes these differences developed and subsequently were maintained.

Here, we assembled two giant panda reference genomes of the QLI and SC subspecies through state-of-the-art whole genome sequencing, including single-molecule sequencing, 10X Genomics, Hi-C, and Chicago HiRise technologies. The combination of these sequencing techniques provided updated genomes of much higher quality, superior contiguity, and accuracy than the existing giant panda genomes. Moreover, together with whole genome deep resequencing, we recalculated the split time of the QLI and SC subspecies. Finally, we examined genes and pathways possibly linked to the relatively small organ sizes and low reproduction capacity of the giant panda.

2. Materials and methods

2.1. Genome library construction and sequencing

For PacBio sequencing, libraries were constructed with an average insert size of 20 kb using SMRTBell Template Pre-Kits (Pacific Biosciences, Menlo Park, USA). A total of 149 single-molecule real-time (SMRT) cells were run on the PacBio Sequel instrument for the SC subspecies, which yielded 285.95 Gb filtered polymerase reads. For the QLI subspecies, 285 SMRT cells were run and yielding 252.58 Gb data. For 10X sequencing, we generated 226.54 Gb and 297.42 Gb raw data representing for the SC subspecies and QLI subspecies, respectively. Two Hi-C libraries were prepared for each subspecies. The libraries of the SC and QLI subspecies were sequenced using the Illumina HiSeq X Ten platform (Illumina, San Diego, USA) and BGISEQ-500 sequencer (BGI-Shenzhen, Shenzhen, China), respectively. Together, the Hi-C library provided approximately 123× (300.1 Gb) and 247× (580.7 Gb) sequencing data for the SC and QLI subspecies, respectively. Three Chicago libraries were prepared, and the libraries provided approximately 122× (286.9 Gb) sequencing data for the QLI subspecies.

For long-read RNA sequencing (Iso-Seq), lung, kidney, liver, and ovarian tissues were collected from an SC individual. Sequencing using a PacBio Sequel sequencer generated a total of 25.593 Gb Iso-Seq raw data with an N50 of 2303 base pairs (bp). For short-read DNA sequencing, DNA samples were sequenced on the Illumina Sequencer HiSeq X Ten with paired-end 150-bp reads. We generated approximately 427.6 and 144.1 Gb raw sequencing data for the SC and QLI subspecies, respectively. RNA was isolated from the blood samples of QLI individuals for short-read sequencing. In total, 26.41 Gb of RNA-seq data were produced for all samples. For genome resequencing, we sequenced the captive and wild individuals using the Illumina HiSeq X Ten sequencer and BGISEQ-500 sequencer, respectively.

2.2. Genome assembly

The genomes of the SC and QLI subspecies were assembled using the Falcon genome assembler (v0.5) [10]. Sequence contigs were error-corrected using the Quiver consensus-calling algorithm (v2.3.1) [11].

For the SC subspecies, primary contigs from the Falcon assembly and all raw PacBio reads were used for gap filling by PBjelly (v15.8.24) [12]. However, since we did not find significant improvement in contigs after this step, we omitted this step for the QLI subspecies. Next, the polished genome assembly and 10X Genomics reads were used as inputs for the fragScaff software (v1.40324) [13] with the following parameters ‘’-fs1 -m 3000 -q 30 -E 30000 -o 60000 -fs2 -C 2 -fS3 -j 1.5 -u 3’’. Subsequently, for the SC subspecies, approximately 50× coverage of Illumina short reads were used to polish the assembly using Pilon (v1.18) [14]. The scaffolds were used as input for a second round of scaffolding using Hi-C reads. According to the alignment information, the hybrid scaffolds were further assembled into super-scaffolds using LACHESIS (v201701) [15]. For the QLI subspecies, we fed the genome into the Dovetail HiRise scaffolding software (v1.0) [16]. The raw PacBio reads were used for gap filling based on PBjelly and error-corrected using Pilon. Finally, to construct the chromosome-level genome of the QLI subspecies, HiC-Pro was performed. The chromosome-level genome was generated by 3D-DNA (v1800922) [17].

2.3. Gene annotation

Gene prediction were carried out following the procedure described in previous study [18]. Briefly, de novo gene prediction, RNA-seq data and homologous proteins were used to annotate the protein-coding genes for the two assemblies. To generate a comprehensive and non-redundant gene set, we used EvidenceModeler (v1.1.1, EVM) [19] to combine all gene structures obtained from de novo approach, homology-based approach and RNA-Seq-based evidence. To identify untranslated regions (UTRs) and acquire information on alternative splicing, we used PASA2 (v2.3.3) to update the gene models [19]. To obtain a high confidence gene set, we removed some outliers.

2.4. Structural variant detection

To detect structural variants via whole genome alignments, pairwise whole genome alignment between the SC and the QLI assembly was carried out using MUMmer (v3.23) [20] (nucmer --numreconx-noextend). Alignments were filtered using delta-filter tools implemented in MUMmer with parameters ‘’-i 90 -l 100’’. 

We aligned approximately 30× QLI subspecies PacBio reads to the SC assembly using NGMLR (v0.2.6) [21] with default parameters. Structural variants (SVs) were detected using Sniffles software (v1.0.6) [21] with parameters of "-s 5 -q 20 -l 50". To eliminate mapping bias, we mapped 30× SC Pacbio reads to the QIL genome using the same method used for the SC genome. Self-heterozygous SVs that would affect bias were removed from the analysis.

### 2.5. Ultra-conserved noncoding elements (UCNEs)

UCNEs were identified according to the method used for the parrot genome [22]. UCNE consensus sequences from human (n = 4351) were downloaded from UCNEBase (https://ccg.epfl.ch/UCNEbase/). Orthologous UCNEs were identified in our two assembled species and four other species—polar bear (U. maritimus), dog (C. familiaris), ferret (M. putorius furo), red panda (A. fulgens)—by aligning the human UCNE set to each genome using BLAST (v2.5.0). Parameters included "-dust no -evalue 1e-5 -max_target_seqs 1 -max_hsps 1 -outfmt 6". The missing UCNEs in the giant panda were validated by the alignment of all the read data to the Human UCNEs. To identify divergent UCNEs in the giant panda, we employed phylogenetic ANOVA in the R package GEIGER (v2.0.6; https://github.com/mwpennell/geiger-v2).

### 2.6. Ka/Ks calculations of Hippo pathway genes

Core orthologs in the Hippo pathway of the giant panda (A. melanoleuca), polar bear (U. maritimus), cat (F. catus), dog (C. familiaris), ferret (M. putorius furo), domestic cow (Bos taurus), and human (H. sapiens) were collected and aligned by the PRANK probabilistic multiple alignment program. PAML (v4.8) [23] was used to calculate the Ka/Ks values of each gene under the free-ratio model (model = 1) and the null model (model = 0). Genes with a significantly faster rate for the foreground branch than the background branches and likelihood ratio test (LRTs) P < 0.5, indicated fast-evolving genes.

### 2.7. Convergent evolution analysis

Convergent evolution analysis was carried out using a previously described pipeline [24]. Single-copy orthologs for the nine selected species—giant panda (A. melanoleuca), polar bear (U. maritimus), cat (F. catus), dog (C. familiaris), ferret (M. putorius furo), red panda (A. fulgens), cheetah (A. jubatus), tiger (P. tigris), and human (H. sapiens)—were identified by OrthoMCL (v2.0) [25] and 6208 single-copy orthologs were identified. Each single-copy ortholog family gene was aligned by PRANK (v100802; https://www.ebi.ac.uk/goldman-srv/webprank/) [26] and then fed into PAML (v4.8). According to the pipeline, we used the JTT-f gene model to identify convergent genes. To exclude orthologs having a discordant gene tree, which would cause false results, we removed genes for which the gene tree was inconsistent with the species tree. Finally, we obtained genes with convergent signatures.

### 2.8. Variants calling and annotation

The mapping was performed using Burrows-Wheeler aligner (bwa, 0.7.17-r1198-dirty) [27] with the SC subspecies as the reference genome. We performed variant calling using GATK (v4.0.3.0) [28] with the function HaplotypetCaller to generate a genome variant call format (gvcf) file for each individual. Joint calling was then performed to generate the combined raw variant call format (vcf) file. Hard filtering was applied to the raw variant set using “QD < 2.0 || FS greater than 60.0 || MQ < 40.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0 –filter-name snp_filter”

### 2.9. Population differentiation (FST), principal component analysis (PCA), ancestral structure analysis, and phylogenetic tree

We used all bi-allelic single-nucleotide polymorphism (SNP) sites to calculate the pairwise Weir and Cockerham's FST [29] between each pair of the three populations (QLI, MSH, and QLA) using VCFTools (v0.1.13) [30]. Genome-wide Complex Trait Analy-
sis (GCTA, v1.92.2 beta) [31] was used to perform the PCA analysis and display the relationship among the three giant panda populations. We carried out an unsupervised estimation of ancestral components for each of the 25 individuals using ADMIXTURE (v1.30) [32] with $K$ values ranging from 2 to 5. The phylogenetic tree was constructed using PHYLIPI (v3.698; https://evolution.genetics.washington.edu/phylip/phylip.html) with 100 bootstraps.

2.10. Population demographic history inference and split time between QLI and SC subspecies

We first used MSMC2 (v2.1.2) [33] to infer the divergence time of the QLI and SC populations. We randomly selected two individuals from each of the two populations involved in the comparison. The following parameters were used: --skipAmbiguous -I 0-4,0-5,0-6,0-7,1-4,1-5,1-6,1-7,2-4,2-5,2-6,2-7,3-4,3-5,3-6,3-7 -I 20 -t 6 -p '1^2 + 15^1 + 1^2'. Because MSMC2 needs a phased SNP set for demography inference, we phased all SNPs of each individual by using beagle (v5.0) with default parameters. To validate the results calculated by MSMC2, we further performed the SMC++ (v1.5.2) [34] to infer the split time of the two subpopulations. All 25 individuals were used in the SMC++ analysis. The generation interval and the $\mu$ used here were also 12 years and $1.29 \times 10^{-6}$ [9].

2.11. Nature selection

We used XP-EHH (v20090727) [35] to detect genomic regions exhibiting signatures of positive selection in the SC or QLI subspecies. The raw XP-EHH values were normalized by subtracting the mean XP-EHH value calculated from the whole genome and dividing it by the standard deviation. The top 0.1% of XP-EHH scores were selected for the candidate sites. We then searched for genes in the 5-kb flanking region from both sides of the candidate SNPs and calculated the accumulated XP-EHH scores by adding each XP-EHH score of the top 0.1% SNPs.

2.12. ABBA-BABA test and F3 statistics

We selected the polar bear (U. maritimus) as an outgroup to perform D statistics. The ordered set was (MSH, QLA, QLI, and Polar Bear). We used popstats (preliminary release) [36] to calculate the D value and Z scores using all individuals of each population. To further test if MSH, QLA, or QLI is an admixed population between the other two populations, we performed the F3 statistics using the qp3Pop (v4.35) implemented in Admixtools (v4.1) [37].

2.13. Ethics approval and consent to participate

Animal care and experiments were conducted under the guidelines established by the Regulations for the Administration of Affairs Concerning Experimental Animals (Ministry of Science and Technology, China, 2013) and were approved by the institutional review board of BGI, China (FT19028-2).

3. Results

3.1. Improved de novo assembly of SC and QLI giant panda genomes

Two male adult giant pandas from the SC and the QLI subspecies were selected for chromosome-level genome assembly (Fig. 1a, b). The sequencing technologies included SMRT sequencing (Pacific Bioscience), short-read (paired-end) sequencing (Illumina), 10X genomics (10X Genomics), and Chicago HiRise and Hi-C technology (Table S1a, b online). More than 100× PacBio sub-reads were first error-corrected and de novo assembled into primary contigs (Fig. S1a, b online). This resulted in contig N50 values of 10.8 Mb for the SC subspecies and 10.5 Mb for the QLI subspecies (Table S2a, b online). We next scaffolded these two genomes with 10X linked reads, which yielded a scaffold N50 of 14.37 Mb for the SC subspecies and 22.66 Mb for the QLI subspecies. Lastly, the Hi-C technology was employed to order and orient the scaffolds into chromosome-scale genomes (Fig. S2a, b online). The final assemblies yielded a scaffold N50 and total length of 136.6 Mb and 2432 Mb for the SC subspecies, and 132.9 Mb and 2352 Mb for the QLI subspecies, respectively (Table 1). Compared with the estimated genome size of the SC (2.44 Gb) and QLI subspecies (2.45 Gb), the assembly completeness ratio for the SC and QLI subspecies was 99.60% and 95.91%, respectively (Fig. S3 online). The giant panda genome contains 21 pairs of chromosomes (2n = 42), including 20 pairs of autosomes and one pair of sex chromosomes. More than 95% of the sequenced bases were assigned to 21 chromosomes (Fig. 1c, and S4a, b online). Compared to the previously published short-read assembled giant panda genome, our newly assembled genome of the SC subspecies showed 106.7- and 5.8-fold improvements in scaffold contiguity over that of the AilMe1.0 [7] and ASM200744v2 [8], respectively (Fig. S5 online). Alignment of our SC subspecies assembly to the ASM200744v2 closed 144.6 Mb gaps in length, accounting for 5.9% of the ASM200744v2 genome (Fig. S6 and S3 online). Further analysis of these closed gap sequences revealed that approximately 63.1% consisted of various classes of long interspersed nuclear elements, short interspersed nuclear elements, and long terminal repeats (LTRs) (Fig. S7 online). Notably, 2707 Mb of the filled gaps map within coding regions, adding 1970.4 kb of protein-encoding sequences (Table S3 online). We also performed synteny analysis by aligning our SC subspecies genome to the ASM200744v2 genome at the chromosome level. This revealed a large proportion of synteny blocks across the whole genome, which accounted for 95.12% and 94.68% of the SC subspecies genome and the ASM200744v2 genome, respectively (Fig. S8 online).

To further assess the completeness of the two giant panda assemblies, we downloaded eight previously published BAC sequences of the SC subspecies from GenBank and aligned them to our two newly assembled genomes. Each BAC aligned uniquely and up to 99% of the BAC regions were covered by the assembled genome, demonstrating a high consensus at the sequence level (Fig. S9a, b, and Table S4a, b online). The assembled genomes displayed high integrity, with over 99.56% of the two genomes supported by paired-end sequence data of the short insert size (Table S5 online). Benchmarking universal single-copy orthologs (BUSCO) analysis was performed to evaluate the completeness of our assembled genomes [38]. The vast majority of the mammalian orthologs could be found in the SC (95.9%) and QLI subspecies (95.7%), indicating the high completeness of these two assemblies (Table S6 online), which exceeded those of the previously assembled giant panda genomes.

3.2. Improved genome annotation

Repetitive elements are major components of eukaryotic genomes and are widely dispersed throughout the genome. We identified a total of 1081.48 and 862.25 Mb of repetitive sequences, accounting for 44.49% and 36.72% of the total genome size for SC and QLI subspecies, respectively (Fig. 1c, Fig. S10a, b, and Table S7 online). The different content of repeats may reflect the slightly smaller assembled genome of the QLI subspecies and/or an expansion of repeats in the SC subspecies. The repeat content of our SC genome was 3.0% and 9.8% higher than that of ASM200744v2 and AilMe1.0 [7] (Table S8 online). This result reflects the advantage of long-read sequencing technology for assembling highly repetitive regions.
We utilized long-read transcriptome sequencing and short-read transcript data to support gene annotation. We obtained 26 Gb of Illumina RNA-seq data and generated 51,181 full-length high-quality transcripts for the SC subspecies. We annotated the protein-coding genes by combining evidence from transcript mapping, homology-based protein mapping, and \textit{ab initio} predictions. Finally, we predicted with high confidence a total of 22,095 and 23,759 protein-coding genes for the SC and QLI subspecies, respectively (Fig. 1c and Table S9a, b online). The values were 5.3% and 11.3% higher than that of AilMel\_1.0, respectively, again demonstrating the superiority of chromosome-level genome assemblies. Among these genes, over 86% were supported by transcript data (87.5% for SC and 86.26% for QLI). BUSCO evaluation showed that the completeness of gene sets was 96.6% for the SC giant panda and 94.5% for the QLI giant panda (Table S10 online). The distribution of gene features is consistent with previous results (Fig. S11 online). Finally, more than 91.5% of the genes in the two gene sets were functionally annotated (Table S11 online). A comprehensive comparison with our newly assembled SC subspecies harbored 742 more genes. GO enrichment analysis determined that these genes were enriched in nucleosome assembly (GO: 0006334, \(P = 1.4 \times 10^{-5}\)), ribonucleoside-diphosphate reductase activity (GO: 0004748, \(P = 2.8 \times 10^{-4}\)), and chromatin organization (GO: 0006325, \(P = 1.4 \times 10^{-4}\)) (Table S12 online). Finally, annotation of noncoding RNA genes of the two subspecies revealed similar copy numbers for each genome (Table S13 online).

3.3. Global genome comparison of SC and QLI subspecies

We aligned the genome of the QLI subspecies to the SC subspecies. Approximately 99.46% of the QLI genome matched in one-to-one syntenic blocks with 96.20% of the SC genome (Fig. 1d). This result demonstrated the high concordance between the two genomes.

As SVs can affect gene expression, phenotype, and even the survival of animals, we further performed genome-wide detection of deletions (InDels) were identified by comparing the two subspecies genomes. We annotated the protein-coding genes by combining evidence from transcript mapping, homology-based protein mapping, and \textit{ab initio} predictions. Finally, we predicted with high confidence a total of 22,095 and 23,759 protein-coding genes for the SC and QLI subspecies, respectively (Fig. 1c and Table S9a, b online). The values were 5.3% and 11.3% higher than that of AilMel\_1.0, respectively, again demonstrating the superiority of chromosome-level genome assemblies. Among these genes, over 86% were supported by transcript data (87.5% for SC and 86.26% for QLI). BUSCO evaluation showed that the completeness of gene sets was 96.6% for the SC giant panda and 94.5% for the QLI giant panda (Table S10 online). The distribution of gene features is consistent with previous results (Fig. S11 online). Finally, more than 91.5% of the genes in the two gene sets were functionally annotated (Table S11 online). A comprehensive comparison with our newly assembled SC subspecies harbored 742 more genes. GO enrichment analysis determined that these genes were enriched in nucleosome assembly (GO: 0006334, \(P = 1.4 \times 10^{-5}\)), ribonucleoside-diphosphate reductase activity (GO: 0004748, \(P = 2.8 \times 10^{-4}\)), and chromatin organization (GO: 0006325, \(P = 1.4 \times 10^{-4}\)) (Table S12 online). Finally, annotation of noncoding RNA genes of the two subspecies revealed similar copy numbers for each genome (Table S13 online).

3.4. Phylogeny, genetic relationship, admixture, and genetic diversity

We further collected DNA from 25 giant panda samples of the SC subspecies (comprising populations from the SC-MSH and SC-QLA) and the QLI population to examine genetic relationships based on our improved assembled genomes (Fig. 2a and Tables S15, S16 online). Of these 25 samples, 19 were obtained from skin samples of animals preserved several decades ago. To evaluate the DNA damage pattern of these old samples, we performed DNA damage analysis by calculating the frequency of C-to-T and G-to-A transitions caused by deamination at the 5’ and 3’ ends of the DNA fragments. Of note, we did not detect elevated deamination-induced C-to-T and/or G-to-A changes at either end of the reads (Fig. S13 online). Finally, we identified 1.042 \(\times\) 10⁷ SNPs and 1.50 \(\times\) 10⁵ InDels for downstream population genomics analysis.

Although these three populations displayed extremely similar genetic diversity (Fig. S14 and Table S17 online), phylogenetic tree analysis, PCA, and population structure analysis consistently supported three distinct groups corresponding to the QLI, SC-MSH, and SC-QLA populations (Fig. 2b, c). The population structure and tree were also validated by including published whole genome sequencing (WGS) data [9] (Figs. S15, 16 online). In the admixture analysis, the QLI, SC-MSH, and SC-QLA populations were divided into three distinct populations when \(K = 3\) (Fig. 2d and Fig. S17 online). This was corroborated by cross-validation and was consistent with phylogeny and biogeographical distributions, indicating limited gene flow between the QLI and SC subspecies. Weir and Cockerham’s fixation index (\(F_{ST}\)) analysis showed a large \(F_{ST}\) value of 0.14 between the QLI and SC subspecies, which was much larger than that between African and Asian human populations [39] (Fig. S18 online), suggesting a considerable genetic distance between the QLI and SC subspecies. Our F3 statistics indicated that none of the three populations were admixed by the other two populations (Fig. 2e). The ABBA-BABA test indicated more shared-derived alleles between the QLI and SC-MSH populations (Polar Bear, QLI, SC-MSH, and SC-QLA, \(D = -0.055, Z = -18.70\)), which was consistent with the \(F_{ST}\) results and geographical distribution.

3.5. Population divergence between QLI and SC subspecies

Considering the dispute over the time for the separation between the QLI and the SC subspecies [9], we further inferred the genetic separation of these two subspecies using MSMSc2 software with the parameter of relative cross coalescence rate (RCCR). The RCCR estimation showed that the QLI population diverged from the SC-MSH or SC-QLA populations approximately 30–40 thousand years ago (kya) and suggested that the complete separation occurred approximately 12 kya. The split time between the SC-QLA and SC-MSH populations was estimated to be 3–6 kya (Fig. 3a).
To further support the separation time inferred by MSMC2, we calculated the split between QLI and SC subspecies using SMC++ software. As expected, the estimated time of the separation event between the QLI and SC subspecies was approximately 10 kya, and the split between the SC-MSH and SC-QLA populations was estimated to have occurred approximately 6 kya, consistent with the result inferred by MSMC2 (Fig. 3b).

3.6. Genetic basis for the reduced size of the inner organs of giant panda

The organ size of animals is often determined by a complex set of biological and ecological constraints to ensure survival. Organ growth must also be flexible enough to adapt to environmental challenges, such as injury or changing food availability [40]. It was previously reported that the giant panda has reduced organ sizes, including the brain, liver, and kidney, compared to other eutherian mammals (Fig. 4a), which likely reflects an adaptive evolutionary strategy to save energy [2,3]. To explain how genetics might contribute to these features, we investigated a series of genes involved in the Hippo pathway, which are involved in the control of organ size and development [41]. The genes were compared to their orthologs in polar bear, ferret, red panda, dog, cow, and human. The coactivators YAP and TAZ are the major effectors in the Hippo pathway. Phosphorylation of YAP and TAZ prevents their entry into the nucleus and nullifies the expression of genes involved in cell proliferation, survival, and migration. This results in reduced size of many inner organs relative to other eutherian mammals (Fig. 4b) [42].

Selective constraint analysis of 14 core genes in the Hippo pathway showed a higher Ka/Ks ratio for the giant panda than that of other animals (chi-square test, \( P = 2.9 \times 10^{-11} \)), indicating a significantly higher relaxed constraint in the giant panda’s Hippo pathway. We then calculated the Ka/Ks values for each of the 14 core genes (Table S18 online). The Ka/Ks values of five genes (FRMD6, LAT51, MOB1B, KIBRA, and YAP1) were significantly higher in the giant panda than in other species (Wilcoxon rank sum test, \( P = 0.00084 \)). Furthermore, FRMD6, SAV1, and YAP1 harbored specific changes. By scanning for these changes across all 25 individuals, we determined that all of these changes were fixed in the population. Two of these changes (p.Met300Val and p.Asn311Asp) are in the FERM C-terminal PH-like domain of FRMD6 (Fig. 4c, d). It is possible that the introduction of a negatively charged Asp residue may affect the function of the FERM C-terminal PH-like domain. In addition, giant panda-specific changes in the SAV1 and YAP1 genes were evident (Fig. S19a, b online). Mutations in these genes have been associated with enhanced tumor growth and poor prognosis in cancer patients [43]. It remains to be determined if these changes also influence tissue growth in the giant panda.

We also conducted an analysis of UCNEs corresponding to regulatory elements for several genes related to the Hippo signaling pathway. These elements in the TEAD1, SMAD2, SMAD3, and SOX2 genes of the giant panda exhibited signs of divergent evolution (Fig. S20a–e online). The collective evidence pointed to a possible relationship between alterations in genes encoding components of the Hippo signaling pathway and the reduced inner organ sizes in the giant panda.

3.7. Analysis of genes related to reproductive performance in the giant panda

The giant panda has a lower reproduction rate than other Ursidae animals, with only one cub delivered every 2 years [1]. Giant pandas have diminutive penises, which could lead to weakened sperm transport and stimulation of female genitals, with possible detrimental effects on reproduction [44]. In addition, follicular cysts and closed ovaries have also been described in the giant panda, contributing to their limited reproductive potential [45]. To better understand the impaired reproductive capacity, we examined lineage-specific variations in genes and regulatory elements. One UCNE (DACH2-Frodo), which is probably a regulator of the DACH2 gene, was absent in the extant giant pandas.
We further aligned raw resequencing reads of all 25 individuals to the DACH2-Frodo gene of polar bear, confirming that this UCNE was absent in all individuals (Fig. S21 online). DACH2 is located on the X chromosome and is a transcription factor involved in the development of eyes, brain, and limbs in insects and vertebrates. In addition, DACH2 regulates genes required for the development of the Mullerian duct [46]. Changes in the DACH2 gene have been found more frequently in premature ovarian failure patients than in healthy individuals [47]. Of note, supporting the impact of this altered UCNE in the giant panda, we did not detect the expression of the DACH2 gene based on RNA-seq data. The loss of the DACH2-Frodo also occurred in the genome of the red panda, which is also characterized by a low rate of reproduction [48]. We identified two convergent evolutionary genes (SLC26A8 and SPESP1) related to reproduction in the giant panda and red panda. Two convergent sites were identified in these two genes, Q37R in SLC26A8 and E332Q in gene SPESP1. SLC26A8, a testis anion transporter 1 (TAT1), is a male-germ-cell-specific member of the large family of anion transporters, solute-linked carrier 26. Changes in this gene may cause asthenozoospermia in humans [49]. SPESP1 encodes a human alloantigen protein that is involved in sperm-egg binding and fusion. Experiments on mice have demonstrated that SPESP1 is necessary for the production of normal sperm [50]. Interestingly, we found one specific change (p.Ile432Var) in the C2B domain region of the SYT6 gene, which seemed to be fixed in the giant panda population (Fig. 5b and Fig. S22 online). The SYT6 gene is expressed in human sperm cells and is involved in sperm acrosome reaction [51]. This change is predicted to possibly be damaging, with a score of 0.774 (sensitivity: 0.85; specificity: 0.92) estimated by the PolyPhen2 [52]. The finding indicated that this change might affect the function of this gene. Finally, we also found two specific changes (p.Ala276Thr and p.Thr306Ala) potentially affecting the function of HOXA13, involved in urogenital development [53] (Fig. S23 online). The collective results suggested that changes and/or loss of regulatory elements in genes involved in reproduction may at least partially explain the low reproductive capacity of the giant panda.

Compared with the SC subspecies, the average annual birth rate (ABR) of the QLI subspecies was much lower (ABR_{SC} = 0.2; ABR_{QL} = 0.1; P = 0.006) [54].
Extended Haplotype Homozygosity (XP-EHH) [35] analysis by setting the QLI subspecies as the reference population to identify reproduction-related genes that were under positive selection in the SC subspecies (Fig. S24a, b, and Table S19 online). Interestingly, we identified a positively selected gene, IQCD, in the SC subspecies (Fig. 5c). This gene is an acrosome-associated gene that plays important roles in fertilization and acrosome reaction. The expression of IQCD in spermatozoa is significantly lower in patients with a low fertilization rate or total fertilization failure than in individuals with normal fertility [55]. Structural variation analysis identified a 56-bp insertion located in the upstream region of the RABL2A gene in the QLI subspecies. This gene is a spermatogenesis-related gene that plays an essential role in male fertility, sperm intraflagellar transport, and tail assembly [56]. These genes are potential candidates to at least partially explain the higher fertility in the SC subspecies compared with the QLI subspecies.

4. Discussion and conclusion

The advantages of long-read sequencing technology have provided opportunities to dramatically improve the quality of assembled genomes. Here, we assembled the genomes of the QLI and SC giant panda subspecies at the chromosome level and performed the first genome assembly of the QLI subspecies. Our high-depth whole genome resequencing data support the view that the extant SC and QLI populations are two distinct subspecies. We further reestimated the divergence time between the QLI and SC subspecies, and between the SC-MSH and SC-QLA populations.

With the combination of long-read sequencing, 10X genomics, Chicago HiRise, and Hi-C scaffolding technologies, we assembled two chromosome-level genomes. These by far represent the most contiguous, complete, and high-quality reference genomes for the giant panda.
genome was recently reported, this genome was assembled based on short-read sequencing and is incomplete [8]. Our assembled genome showed a contiguity improvement over 200-fold at the contig level, compared to the previous genome assembly. Moreover, the SC subspecies genome closed more than 144 Mb of gaps and increased repeat mapping compared with the latest giant panda genome (ASM200744v2) [8]. Gene annotation was also improved by the inclusion of data from full-length mRNA sequencing. More than 742 genes found in the present study were absent in the previously published assembly (ASM200744v2), which also reflects the completeness of our genome and gene sets.

We discovered numerous SNPs and InDels as well as a large number of SVs that are specific to each of the SC and QLI giant panda genomes. These differences strongly suggest that the QLI and SC giant panda subspecies are two divergent subspecies, as previously defined [4]. Considering the matching result on the split time between the QLI and SC subspecies estimated by MSMC2 and SMC++, we conclude that the divergence between the two subspecies occurred at 10–12 kya, which supports the previous inference [4], but is far more recent than the 300 kya estimated by other authors [9]. At approximately 6 kya, three genetically distinct giant panda populations evolved with the SC population diverging into the SC-MSH and SC-QLA populations. Paleogenomic studies on ancient giant pandas also support that the split time between the QLI and SC subspecies was more recent than 300 kya [57]. The lower depth resequencing might underestimate SNPs in the population in the data of Zhao et al. [9], which might cause false-negatives in detecting the variations and further influence the accuracy of the inferred split times. The divergence of the QLI and SC subspecies seems not to be influenced by human activities, considering the relatively early time period. Although the global climate was warm after the Last Glacial Maximum, a sudden decrease in temperature occurred in the Younger Dryas (YD) period approximately 12,900 to 11,700 years ago [58]. It is possible that the sudden drop in temperature during the YD changed the forest habitats, especially the bamboo vegetation, which in turn may have separated the giant panda populations. Although the temperature and precipitation gradually increased after the YD [59], the precipitation and temperature in Sichuan Province began to decline again from 6000 years ago. The northern people living in the upper reaches of the Yellow River migrated to the northwest of the SC Plateau [60]. We inferred that this early human migration, together with climate change, might be linked to the divergence of the SC-MSH and SC-QLA populations.

Fig. 5. Possible genetic mechanisms of the low reproductive ability of the giant panda. (a) The loss of a regulatory element in the DACH2 gene in the giant panda. (b) Comparative genomic analysis revealing a giant panda-specific mutation in the SYT6 gene. The heat map under the amino sequence alignment indicates the functional effect prediction for the I432V by PolyPhen-2, where a mutation will be predicted as “possibly damaging” if the estimated score is 0.447 to 0.908. (c) The distribution of the normalized XP-EHH scores around the IQCD gene. The gray bar indicates the upstream and downstream 50-kb region of the IQCD gene. The pink bar indicates the IQCD gene region.

As an endangered species, the low reproduction of the giant panda has become an obstacle for population recovery and conservation. We observed that a number of genes associated with reproductive performance, including DACH2, SYT6, HOXA13, SLC26A8,
and SPESP1, exhibited unique evolutionary traits in the giant panda genome. The finding suggests that such features may at least partially underlie the impaired reproductive capacity of the giant panda. Thus, lost UCNE elements of the DACH2 gene likely contribute to the genetic basis of the low reproductive capacity of the giant panda. We identified one fixed change in the SYT6 gene and two fixed changes in the HOXA13 gene in the giant panda genome. These changes may influence the function of these genes, possibly affecting sperm production and male genitals. Interestingly, the HOXA13 gene was not completely assembled in any of the previous giant panda genomes, which supports the need to update the giant panda genome.

In conclusion, our genome assembly of the two extant giant panda subspecies represents an important resource to understand the peculiar phenotypes of the giant panda and to improve breeding and conservation of the giant panda.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key Program (2016YFC0503200) from the Ministry of Science and Technology of China, a special grant for the giant panda from the State Forestry Administration of the People’s Republic of China, the Fundamental Research Funds for the Central Universities of the People’s Republic of China, the Foundation of Key Laboratory of State Forestry and Grassland Administration (State Park Administration) on Conservation Biology of Rare Animals in the Giant Panda National Park (KLSPFAGP2020.002), and the Guangdong Provincial Key Laboratory of Genome Read and Write (2017B030301011). We thank Mr. Wei Wang for his help on painting the figures of giant panda, the Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (BGI-Shenzhen, Guangdong, China), and the China National GeneBank for producing sequencing data. Maps in this article were reviewed by Ministry of Natural Resources of the People’s Republic of China (GS(2021)146).

Author contributions

Sheng-Guo Fang conceived and initiated the project. Yan Huang, Minchun Zhang, Rengui Li, Zhizhong Zhang, Yinghu Lei, Ling Zhang, Heming Zhang, Desheng Li, Maijia Qiao, Daifu Wu, Pengpeng Zhao, Jian-Qing Lin, and Danhui Zhang organized and collected the samples. Rengui Li, Xiaopeng Li, Hong Li, Keyi Tang, Qiqi Liang, Wenkai Jiang, Xun Xu, Sunil Kumar Sahi, and Xin Liu performed DNA extraction, library construction, and sequencing. Xuanmin Guan, Hong Li, Wenkai Jiang, and Qiqi Liang performed the genome assembly. Xuanmin Guan, Tianming Lan, Qiu-Hong Wan, Hong Li, Yan Xu, Minchun Zhang, and Huan Liu, performed comparative genomics analysis. Xuanmin Guan, Tianming Lan, Hai-Meng Li, Xun Xu, Xin Liu, Huanming Yang, and Huan Liu, performed the population genomics analysis. Xuanmin Guan, Tianming Lan, Qiu-Hong Wan, and Huan Liu wrote the manuscript. Michael Lisby, Karsten Kristiansen, and Sheng-Guo Fang extensively revised the manuscript. Karsten Kristiansen, Huan Liu, and Sheng-Guo Fang provided supervision. All authors have read and approved the final manuscript.

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2021.02.002.

References

[19] Wei H, Zhang J, Zhao, Jian-Qing Lin, and Danhui Zhang organized and collected the samples. Rengui Li, Xiaoping Li, Hong Li, Keyi Tang, Qiqi Liang, Wenkai Jiang, Xun Xu, Sunil Kumar Sahi, and Xin Liu performed DNA extraction, library construction, and sequencing. Xuanmin Guan, Hong Li, Wenkai Jiang, and Qiqi Liang performed the genome assembly. Xuanmin Guan, Tianming Lan, Qiu-Hong Wan, Hong Li, Yan Xu, Minchun Zhang, and Huan Liu, performed comparative genomics analysis. Xuanmin Guan, Tianming Lan, Hai-Meng Li, Xun Xu, Xin Liu, Huanming Yang, and Huan Liu, performed the population genomics analysis. Xuanmin Guan, Tianming Lan, Qiu-Hong Wan, and Huan Liu wrote the manuscript. Michael Lisby, Karsten Kristiansen, and Sheng-Guo Fang extensively revised the manuscript. Karsten Kristiansen, Huan Liu, and Sheng-Guo Fang provided supervision. All authors have read and approved the final manuscript.

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2021.02.002.

2011
Xuanmin Guang graduated from Sichuan University with a master’s degree and is a Ph.D. candidate at Zhejiang University under the supervision of Prof. Sheng-Guo Fang. His research interest mainly focuses on genomics and evolution.

Yan Huang is a research fellow at China Conservation and Research Center for the Giant Panda. His research focuses on conservation biology and the breeding of the giant panda.

Hong Li obtained her master’s degree from Shanghai Ocean University in 2016 under the supervision of Yong-sheng Tian. She then joined Novogene Bioinformatics Institute for Life Sciences, Qingdao, China. After graduation she joined the Nangjing Bioinformatics Center. Her research mainly focuses on bioinformatics and bioinformatics.

Karsten Kristiansen is a professor of Molecular Biology and head of the Laboratory of Genomics and Molecular Biomedicine at the Department of Biology, University of Copenhagen. He is also a professor and director of the Institute of Metagenomics at BGI Shenzhen and the Institute of Metagenomics, Qingdao-Europe Advanced Institute for Life Sciences, Qingdao, China. After graduation from the University of Copenhagen, he held research positions at the Max-Planck-Institut für Molekulare Genetik in Berlin and the Institut de Biologie Physico-Chimique, Fondation Edmond de Rothschild in Paris. His research focuses on interactions between host genome, host immune system, and gut microbiota.
Huan Liu is a research fellow at BGI-Shenzhen and head of the Institute of Digital Earth. He is also the Deputy Director of the State Key Laboratory of Agricultural Genomics and the leader of the National Key Research and Development Project. He joined BGI in 2010, focusing on genomic and phylogenomics research of animals and plants.

Sheng-Guo Fang is a full professor at the College of Life Science, Zhejiang University, and the director of the State Conservation Center for Gene Resources of Endangered Wildlife. His research focuses on the conservation biology of endangered animals, molecular ecology, and conservation genomics.