The sequence and de novo assembly of the giant panda genome

Ruiqiang Li1,2, Wei Fan1, Geng Tian1,3, Hongmei Zhu1, Lin He4,5, Jing Cai3,6, Quanfei Huang1, Qingle Cai1,7, Bo Li1, Yinqi Bai1, Zhihe Zhang1, Yaping Zhang2, Wen Wang6, Jun Li1, Fuwen Wei1, Heng Li1, Min Jian1, Jianwen Li1, Zhaolei Zhang1, Rasmus Nielsen12, Dawei Li1, Wanjun Gu13, Zhentao Yang1, Zhaoling Xuan1, Oliver A. Ryder14, Frederick Chi-Ching Leung15, Yan Zhou1, Jianjun Cao1, Xiao Sun16, Yonggui Fu17, Xiaodong Fang1, Xiaosen Guo1, Bo Wang1, Rong Hou8, Fujun Shen8, Bo Mu1, Peixiang Ni1, Runmao Lin1, Wubin Qian1, Guodong Wang1,3,6, Chang Yu1, Wenhui Nie6, Jinhuan Wang6, Zhigang Wu1, Huiqing Liang1, Jiuemeng Min1,2, Qi Wu9, Shifeng Cheng2,6, Jue Ruan1,3, Mingwei Wang1, Zhongbin Shi1, Ming Wen1, Binghang Liu1, Xiaoli Ren1, Huisong Zheng1, Dong Dong1,11, Kathleen Cook11, Gao Shan1, Hao Zhang1, Carolin Kosiol18, Xueming Xie13, Zhuhong Lu11, Hancheng Zheng1, Yinfu Li1,3, Cynthia C. Steiner14, Tommy Tsan-Yuk Lam15, Siyuan Lin1, Qinghui Zhang1, Guoqing Li1, Jing Tian1, Ming Gong1, Hongde Liu16, Dejin Zhang16, Lin Fang1, Chen Ye1, Yuanbin Zhang1, Wenbo Hu17, Anlong Xu17, Yuanyuan Ren1, Guojie Zhang1,3,6, Michael W. Bruford19, Qibin Li1,3, Liija Ma1,3, Yiran Guo1,3, Na An1, Yujie Hu1,3, Yang Zheng1,3, Yongsheng Li1, Zhiquan Li3, Ping Liu1, Yanling Chen1, Jing Zhao1, Ning Qiu1,7, Shancen Zhao1, Feng Tian1, Xiaolin Wang1, Haiyin Wang1, Lizi Xu1, Xiao Liu1, Tommaso Vinar20, Yajun Wang21, Tak Wah Lam22, Siu-Ming Yiu22, Shiping Liu23, Hemin Zhang24, Desheng Li24, Yan Huang24, Xia Wang1, Guohua Yang1, Zhan Jiang1, Junyi Wang1, Nan Qin1, Li Li1, Jingxiang Li1, Lars Bolund1, Karsten Kristiansen1,7, Gane Ka-Shu Wong23, Maynard Olson26, Xiuqing Zhang1, Songgang Li1, Huanming Yang1, Jian Wang1 & Jun Wang1,2

Using next-generation sequencing technology alone, we have successfully generated and assembled a draft sequence of the giant panda genome. The assembled contigs (2.25 gigabases (Gb)) cover approximately 94% of the whole genome, and the remaining gaps (0.05 Gb) seem to contain carnivore-specific repeats and tandem repeats. Comparisons with the dog and human showed that the panda genome has a lower divergence rate. The assessment of panda genes potentially underlying some of its unique traits indicated that its bamboo diet might be more dependent on its gut microbiome than its own genetic composition. We also identified more than 2.7 million heterozygous single nucleotide polymorphisms in the diploid genome. Our data and analyses provide a foundation for promoting mammalian genetic research, and demonstrate the feasibility for using next-generation sequencing technologies for accurate, cost-effective and rapid de novo assembly of large eukaryotic genomes.

The giant panda, *Ailuropoda melanoleuca*, is at high risk of extinction because of human population expansion and destruction of its habitat. The latest molecular census of its population size, using faecal samples from 3,000 individuals, which were confined to several small mountain habitats in Western China1. The giant panda has several unusual biological and behavioural traits, including a famously restricted diet, primarily made up of bamboo, and a very low fecundity rate. Moreover, the panda holds a unique place in evolution, and there has been continuing controversy about its phylogenetic position2. At present, there is very little genetic information for the panda, which is an essential tool for detailed understanding of the biology of this organism. A major limitation in obtaining extensive genetic data is the prohibitive costs associated with sequencing and assembling large eukaryotic genomes.

The giant panda, *Ailuropoda melanoleuca*, is at high risk of extinction because of human population expansion and destruction of its habitat. The latest molecular census of its population size, using faecal samples from 3,000 individuals, which were confined to several small mountain habitats in Western China. The giant panda has several unusual biological and behavioural traits, including a famously restricted diet, primarily made up of bamboo, and a very low fecundity rate. Moreover, the panda holds a unique place in evolution, and there has been continuing controversy about its phylogenetic position. At present, there is very little genetic information for the panda, which is an essential tool for detailed understanding of the biology of this organism. A major limitation in obtaining extensive genetic data is the prohibitive costs associated with sequencing and assembling large eukaryotic genomes.
libraries of animal and plant genomes in an efficient and cost-effective way. This opens new possibilities for promoting the construction of reference sequences for other species. For example, the assembly of the giant panda genome will have far-reaching implications for this endangered species. Furthermore, our demonstration that the panda genome contains numerous repeat sequences, including tandem repeats and transposable elements, and contributes to disease control and conservation efforts. This work presents a powerful tool for investigating the genetic basis of panda's unique biology and for contributing to conservation efforts.

**Short-read de novo sequencing and assembly**

For sequencing, we selected a 3-year-old female giant panda from the Chengdu breeding centre in China. The panda genome contains 20 pairs of autosomes and one pair of sex chromosomes (2n = 42) (Supplementary Fig. 1). We used a whole-genome shotgun sequencing strategy and Illumina Genome Analyser sequencing technology. DNA was extracted from the peripheral venous blood, and 37 paired-end sequencing libraries were constructed with insert sizes of about 150 base pairs (bp), 500 bp, 2 kb, 5 kb and 10 kb. In total, we generated 176-Gb of usable sequence (equal to 73-fold coverage of the whole genome), with an average read length of 52 bp (Supplementary Tables 1 and 2).

We assembled the short reads using SOAPdenovo (http://soap.genomics.org.cn)—a genome assembler developed specifically for use with next-generation short-read sequences (Supplementary Fig. 2). SOAPdenovo uses de Bruijn graph algorithm and applies a stepwise strategy to make it feasible to assemble the genome using a supercomputer (32 cores and 512 Gb random access memory (RAM)). The algorithm is sensitive to sequencing errors, so we excluded the data generated from poor libraries, filtered low-quality reads, and used the 134 Gb (56-fold coverage) high-quality reads for de novo assembly.

We first assembled the short reads from fragmented small insert-size libraries (<500 bp) into contigs using sequence overlap information. Contigs were not extended into regions in which repeat sequences created ambiguous connections. At this point, we assembled about 39-fold coverage short-reads into contigs having an N50 length of 1.5 kb, achieving a total length of 2.0 Gb (Table 1). Here, we avoided using reads from long insert-size paired-end libraries (≥2 kb) on contig assembly because these libraries were constructed using a circularization and random fragmentation method, and the small fraction (~5%) of chimeric reads in these long insert-size libraries could generate incorrect sequence overlap resulting in misassembly.

We then used the paired-end information, step by step from the shortest (150 bp) to the longest (10 kb) insert size, to join the contigs into scaffolds. We obtained a scaffold N50 length of 1.3 Mb and a total length of 2.3 Gb, determined by counting the estimated intra-scaffold gaps. Most of the remaining gaps probably occur in repetitive regions, so we further gathered the paired-end reads with one end mapped on the unique contig and the other end located in the gap region and performed local assembly with the unmapped end to fill in the small gaps within the scaffolds. The resulting assembly had a final contig N50 length of 40 kb (Table 1). In total, 223.7 Mb gaps were closed. Roughly 54.2 Mb (2.4% of total scaffold sequence) remained unsealed, of which we determined that about 90% contained carnivore-specific transposable elements and the remainder were primarily tandem repeats with high unit identity and lengths larger than the sequencing read length, which could not be assembled with the current data. About 0.05% of the panda assembly was composed of tandem repeats (Supplementary Table 3). The genome similarity between the panda and the dog, and that 0.2% of the dog genome is made up of tandem repeats, we estimate that around 3.6 Mb (0.15%) of tandem repeat sequences might be missing in the current panda genome assembly.

To evaluate the single-base accuracy of the assembled genome sequence, we realigned all the usable sequencing reads onto the scaffolds using SOAPaligner. The peak sequencing depth was 65×, and more than 20 reads covered over 99% of the assembled sequences (Fig. 1a). Using massively parallel next-generation sequencing technology and improving sequencing depth to more than 20× coverage provides a very high single-base accuracy.

The GC content difference is a primary factor for non-random sequencing-depth distribution. Using 500-bp non-overlapping sliding windows along the genome, which is similar to the size of sequencing DNA fragments (150–500 bp), we found that both very low and very high GC regions had a relatively lower sequencing depth, but nearly all regions with a GC content between 20% and 80% had more than 20× coverage, which is sufficient for de novo assembly (Supplementary Fig. 3). Only a minor fraction (0.004%) of the genome—compared to the sequenced dog (0.079%), human (0.095%) and mouse (0.015%) genomes—had a GC content lower than 20% or higher than 80% (Supplementary Fig. 4). Assuming similar GC content in the panda and dog genomes, 0.075% (1.8 Mb) sequence with extreme GC content could be missing in the current panda assembly. Nevertheless, the analysis showed that de novo genome assembly is unlikely to be strongly affected by GC-biased non-random sampling.

**Table 1** Summary of the panda genome sequencing and assembly

<table>
<thead>
<tr>
<th>Step</th>
<th>Paired-end insert size (bp)*</th>
<th>Sequence coverage (&gt;×)†</th>
<th>Physical coverage (&gt;×)‡</th>
<th>N50 (bp)</th>
<th>N90 (bp)</th>
<th>Total length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial contig</td>
<td>110–230, 380–570</td>
<td>38.5</td>
<td>96</td>
<td>1,483</td>
<td>224</td>
<td>2,021,639,596</td>
</tr>
<tr>
<td>Scaffold 1</td>
<td>Add 1,700–2,800</td>
<td>8.4</td>
<td>151</td>
<td>32,648</td>
<td>7,780</td>
<td>2,213,848,409</td>
</tr>
<tr>
<td>Scaffold 2</td>
<td>Add 3,700–7,500</td>
<td>6.5</td>
<td>450</td>
<td>229,150</td>
<td>45,240</td>
<td>2,250,442,210</td>
</tr>
<tr>
<td>Scaffold 3</td>
<td>Add 9,200–12,300</td>
<td>56.0</td>
<td>373</td>
<td>581,933</td>
<td>127,336</td>
<td>2,297,100,301</td>
</tr>
<tr>
<td>Scaffold 4</td>
<td>Add 11,000–13,000</td>
<td>84.0</td>
<td>1,070</td>
<td>1,281,781</td>
<td>312,670</td>
<td>2,299,498,912</td>
</tr>
<tr>
<td>Final contig</td>
<td>All</td>
<td>39.86</td>
<td>9,848</td>
<td>2,245,302,481</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Add denotes accumulative; for example, scaffold 2 uses data of 110–230, 380–570 and 1,700–2,800.

*Approximate average insert size of Illumina Genome Analyser sequencing libraries. The sizes were estimated by mapping the reads onto the assembled genome sequences.

†High-quality read sequences that were used in assembly. Coverage was estimated assuming a genome size of 2.4 Gb. Sequence coverage refers to the total length of generated reads, and physical coverage refers to the total length of sequenced clones of the libraries.

‡N50 size of contigs or scaffolds was calculated by ordering all sequences then adding the lengths from longest to shortest until the summed length exceeded 50% of the total length of all sequences. N90 is similarly defined.
There are 27 known panda mRNA genes in GenBank, one of which is the SRY sex determination gene located on chromosome Y, thus not present in the female panda. We were able to detect the remaining 26 genes in the assembled scaffolds with 99.3% total sequence alignment (Supplementary Table 4). All of these genes, with the exception of the ribosomal gene RPS15, were intact on the scaffolds. For RPS15, we found one fragment in the scaffolds and the other in a small contig not incorporated in the scaffolds. These data indicate that the genome assembly has good coverage and completeness for genes with unique sequences, but given their nature, multiple-copy genes or genes containing repetitive sequences may be fragmented.

To assess the large-scale and local assembly accuracy of the scaffolds, we sequenced and assembled nine bacterial artificial chromosomes (BACs) independently using Sanger sequencing technology. Each BAC (average length 97.7 kb) was aligned to only one scaffold, and up to 98.0% of the total BAC regions were well covered by the assembled contigs. We did not observe any obvious misassembly errors (Fig. 1b and Supplementary Fig. 5). We then checked the single-base mismatch or small insertion/deletion differences. After excluding the annotated heterozygous single nucleotide polymorphisms (SNPs) and indels (described later), the rate of single-base differences was 0.07% and the average number of insertions/deletions on each BAC was 16 (Supplementary Table 5 and Supplementary Fig. 6). There was both a very high read depth on the scaffolds and a high Phred score on the BACs at the discrepant bases, so most of these might be unidentified heterozygous SNPs. Manual inspection showed that 38% of the insertions/deletions were caused by assembly errors on the BAC (Supplementary Fig. 7a), and the rest were unidentified heterozygous indels of the diploid genome (Supplementary Fig. 7b).

To assess genome coverage of the assembled contigs and scaffolds, we first estimated the genome size of the panda. To obtain the best estimate, we used sequencing depth, size ratio of syntenic blocks, and comparison of C-values. Using 17-base oligonucleotides and their frequency distribution in the usable sequencing reads to determine sequencing depth, we obtained an estimated panda genome size of 2.46 Gb (Supplementary Fig. 8). Given the presence of sequencing errors, we expect that the sequencing depth is underestimated, and consequently the panda genome size should be slightly smaller than that of the dog. On the basis of previous reports showing that the karyotype of panda is similar to that of bears, the bear and the panda are expected to have similar genome sizes. We again determined the panda genome size to be 2.40 Gb. Using this genome size, the assembled contigs and scaffolds should cover about 94% and 96% of the whole genome, respectively.

**Panda, dog and human repeat comparison**

Combining results from analyses using the Repbase library, alignment with known transposable-element-related genes, and de novo RepeatModeller identification, we estimate that transposable elements comprised approximately 36.2% of the panda genome, which is similar to that of the dog genome (36.1%), and lower than the human genome (46.1%) (see Supplementary Tables 6 and 7). Analysis of the divergence rate of the transposable element in the panda genome using Repbase transposable-element libraries showed that nearly all of the identified panda transposable-element copies had a >10% divergence rate from the consensus. This high divergence rate may be related to the fact that the Repbase transposable-element consensus sequences were annotated using mammalian genomes other than the panda. Using RepeatModeller transposable elements, we found that about 70 Mb of transposable-element sequences (3% of the genome) had a <10% divergence rate from the consensus (Supplementary Fig. 9), which are likely to be active transposable elements of recent origin. (See Supplementary Results for further analyses on specific transposable elements.)

**Panda genome has a low divergence rate**

To investigate the rate of recent segmental duplication (>90% identity, >1 kb length) in the panda genome, we conducted self-sequence alignment and identified 3,095 duplicated fragments with a total length 10.4 Mb (0.43%) in the whole genome assembly. To determine whether these highly similar recent duplicates were collapsed by the assembly, we aligned all sequencing reads to the assembled genome, and identified 5,485 segments (>1 kb in length) with a total length of 13.9 Mb that had an obviously higher read depth than the other genomic regions. Considering the average depth was about 2.47 times that of the whole genome, we estimated that the total length of the duplicated copies was about 34.3 Mb (1.43% of the whole genome) (Supplementary Table 8). Integrating the results of the two methods, we estimated that recent segmental duplication comprised about 43.7 Mb (1.82%) of the panda genome, which is lower than the rate in the dog genome (106.6 Mb, 4.21%)<sup>15</sup>.

To investigate panda genome conservation and evolution, we performed whole-genome alignment with the panda, dog and human genomes. Each of these genomes contained ~1.4 Gb of non-repeative sequence, of which 846 Mb of sequence was shared by all three genomes. Of the remainder, 322 Mb of sequence was shared between panda and dog, which was much higher than that between panda and human (163 Mb) or between dog and human (58 Mb). The panda
had the lowest amount of lineage-specific sequences, indicating that it has the lowest divergence rate among the three genomes (Fig. 2a).

We found that the panda, dog and human genomes had high genomic synteny. For example, the syntenic relationship of 35 panda scaffolds, 36.5 Mb of human chromosome 2, and dog chromosome 37 (Fig. 2b) showed no large-scale rearrangement within the chromosomal region. For the whole genome, regions of conserved synteny between the panda and dog genomes covered 2.22 Gb (96.7%) and 2.27 Gb (92.9%), respectively and the pairwise syntenic regions between the panda and human genomes covered 2.19 Gb (95.3%) and 2.66 Gb (88.1%), respectively (Supplementary Table 9).

We further detected genomic rearrangement events using the whole genome alignments. Here, we used 3,805 panda scaffolds with lengths longer than 10 kb, which covered 98.6% of the assembled genome. To define rearrangements, we used different cutoffs for minimum syntenic segments. The total intra-chromosomal breakpoints between the panda and the dog genomes varied from 468 with a 5-kb cutoff to 20 with a 100-kb cutoff. We also identified about 42 inter-chromosomal rearrangements using a 100-kb cutoff (Supplementary Table 10). We used the human genome as the common outgroup and found that there were 4–5 times more rearrangements in dog than in panda, which provided evidence that the panda has a lower divergence rate than the dog.

**Gene prediction and comparison**

To predict the number of genes in the panda genome, we used both evidence-based and ab inito methods. We aligned 20,001 (94%) of the human and 19,086 (99%) of the dog genes to the panda genome and predicted 19,303 and 19,245 gene loci, respectively (see Methods and Supplementary Table 11). We also used Genscan and Augustus with model parameters trained on the human, and predicted 44,428 and 29,238 gene loci, respectively for the two programs, in the panda genome. To facilitate further analysis, we integrated all the gene sources and created a reference gene set that contained 21,001 genes for the panda (Supplementary Table 12).

To measure the quality of gene prediction, we compared the length distribution of genes, coding sequences (CDS), exons and introns, and the distribution of exon number per gene among representative mammalian genomes, including panda, dog, human, mouse and opossum. The panda was similar to the human with respect to all of these key parameters (Supplementary Fig. 10). To evaluate the rate of missing exons in the predicted genes, we aligned both the panda and dog genes to the human genes and calculated the percentage of human gene sequences covered. On average, the dog and panda genes covered 96.2% and 93.5% of the human gene sequences, respectively (Supplementary Fig. 11a). Further checking showed that most (more than 80%) of the unannotated exons were at the 5’ or 3’ ends of genes; these exons were usually very small and separated by large introns that were hard to predict by gene finders, although they were correctly presented in the assembled scaffolds (Supplementary Fig. 11b). There are 626 genes, with predicted functions in defence/immunity, receptors and signalling, located in the segmental duplication regions, which was similar to the findings in a previous analysis of the dog genome (Supplementary Table 13). Overall, we found that the quality of the predicted panda genes was comparable to that of other well-annotated mammalian genomes.

We next used InParanoid and Multiparamoid to identify gene orthologues among panda, dog, mouse and human. In total, we found 18,643 orthologous clusters that were shared by at least two species, and 15,060 clusters that were shared by all four species (Supplementary Fig. 12). There were 2,534 panda-specific genes, which is nearly double the amount of dog-specific genes (1,677). This difference might be due to the fact that only an evidence-based method was used for dog gene prediction and a smaller number of genes was predicted in the dog genome (19,305).

Gene gains and losses are one of the primary contributors to functional changes. To obtain greater insight into the evolutionary dynamics of the genes, we determined the expansion and contraction of the gene orthologue clusters among these four species. We used a maximum-likelihood method to estimate the orthologue cluster sizes in their common ancestor and each of the current species. Overall in all four genomes, we found a greater amount of gene contraction than expansion (Fig. 3), indicating that loss-of-function might have an important role in the evolution of these genomes.
role in functional evolution. We also saw that the clusters within different species that underwent the most marked changes (expansion or contraction) were those involved in receptor activity (Supplementary Table 14).

We looked for signatures of positive selection using branch-site likelihood ratio tests for 12,424 genes of high confidence that had 1:1 orthologues in the panda and human genome and in at least one of the rat, mouse and dog genomes that had passed a series of rigorous filters for large-scale synteny, alignment quality, conservation of exon–intron structure, and sequence quality. Using three different tests—one specific for the panda lineage, one specific for the dog lineage, and one combining evidence from all five species included in the alignment—we found 134, 94 and 182 positively selected genes (PSGs), respectively, using a conservative 5% false-discovery-rate criterion.

The panda and the dog lineage only share six PSGs. However, Fisher’s exact and Mann–Whitney U tests showed enrichment of several immunity- and defence-related functional categories in both the panda and the dog genomes, which is consistent with the results from previous genome-wide positive selection scans in mammalian genomes (Supplementary Table 15). Gene Ontology classification showed that panda PSGs spanned a diverse range of immune response systems, in particular acute inflammatory and innate immune response, as well as response to wounding (Supplementary Table 16).

Of note, the inclusion of the panda genome in the multiple species analysis showed an enrichment of two new Gene Ontology categories ‘blood circulation and gas exchange activity’ and ‘haemostasis’ as compared to an analysis with six mammalian species (human, chimpanzee, macaque, mouse, rat and dog). In contrast to the various enrichments for Gene Ontology categories for the panda PSGs, only two categories, ‘cytolysis’ and ‘cell–cell adhesions’, were over-represented in the dog PSGs. The latter category includes cadherins (for example, PCDHB10 (P = 2.66 × 10⁻³²), PCDHB8 (P = 1.80 × 10⁻¹⁹), PCDHB17 (P = 3.74 × 10⁻¹⁹) and ME1 (P = 7.10 × 10⁻⁹)).

**Panda-specific characteristics**

To gain insight into some of the traits unique to the panda, we investigated genes that may influence characteristics such as panda diet and fecundity. For our investigation of genes that may be involved in food selection and digestion, we identified genes that encode digestive enzymes protease, amylase, lipase, cellulase, lactase, invertease and maltase in the panda genome, indicating that the panda probably has all the necessary components for a carnivorous digestive system. We did not find any homologues of digestive cellulase genes, including endoglucanase, exoglucanase and beta-glucosidase, indicating that the bamboo diet of the panda is unlikely to be dictated by its own genetic composition, and may instead be more dependent on its gut microbiome.

Taste is also an important factor in the development of dietary habits. The five components of basic taste are: sweetness, saltiness, sourness, bitterness and umami. Bitterness is sensed by the T2R gene family, and we identified 16 T2R genes in the panda genome, which is comparable to 14 T2R genes reported for the dog. Umami is sensed through the TIR gene family. In the panda genome, TIR2 and TIR3 are in an intact form, but TIR1 has become a pseudogene. We found that the two panda TIR1 exons contain frameshift errors, which we confirmed by Sanger sequencing. We also found that the dS/dN (the ratio of the rate of non-synonymous substitutions to the rate of synonymous substitutions) difference for this gene was low between the panda (0.17) and dog (0.13) lineage, indicating a recent death for this gene in the panda lineage (Supplementary Fig. 13). The TIR1/TIR3 heterodimer is known to be a receptor that senses umami, and a previous study indicated that TIR1/TIR3 may be the sole receptor for umami taste. Umami perception occurs through the detection of the carboxylate anion of glutamic acid, which is a naturally occurring amino acid common in meats, cheese, broth, stock and other protein-heavy foods. Thus, the loss-of-function of the TIR1 gene might prevent the panda from expressing a functional umami taste receptor, which may partly explain why the panda diet is primarily herbivorous despite its taxonomic classification as a carnivore.

The giant panda has a low fecundity rate, which is of great consequence given its endangered status. We identified nearly all of the reproducations known to be critical for mammalian gonad development and function. We identified a putative pseudo follicle-stimulating hormone (FSH) β-subunit gene (giant panda-FSHB2), which has high sequence similarity to the giant panda FSH β-subunit gene (giant panda-FSHB1) in the panda genome as well as in several other mammalian species, including cow and horse (Supplementary Fig. 14). At this stage, whether the pseudo FSHB2 gene contributes to the reproductive features of the giant panda remains to be determined.

**Panda has a high heterozygosity rate**

As the panda population is estimated to be ∼2,500 worldwide at present, the issue of genetic variability and evolutionary potential is extremely important because low genetic variability will cause a population decline. A recent molecular analysis of the 655 bp of the mitochondrial control region DNA and ten microsatellite loci indicated that the panda still has high genetic variability. Sequence data for the whole genome to firmly establish this important issue have not been available until now.

Using the assembled panda genome sequence as a reference, we realigned all of the usable sequencing reads with the genome to identify the heterozygous SNPs, small indels, and structural variations between the two sets of homologous chromosomes in the diploid genome. We were able to align about 85% of the usable sequencing reads. Among these, we used the 90% that had unique best locations to detect variation by applying the pipeline developed in the analysis of the first Asian human genome (Fig. 5a). We identified 2.7 million heterozygous SNPs in the panda diploid genome. The estimated heterozygosity rate of the autosomes was 1.35 × 10⁻³ for the whole genome and 0.66 × 10⁻³ for the coding regions (Fig. 5a). This is about 1.95 times higher than the rate estimated for the human genome (0.69 × 10⁻³ and 0.34 × 10⁻³). The heterozygosity rate in the X chromosome is about half that of the autosomes, and the ratio of transition/transversion is 2.1 (Supplementary Table 18). Our finding of a high heterozygosity rate may be due to the mixed genetic background (Liangshan and Minshan sub-populations) of the panda that we sequenced. These results, however, may indicate that the panda species still has high genetic variability, which would support the potential for its successful survival despite its small population size. To determine this, analyses of other panda genomes are required.

The heterozygous SNP rate in the panda genome showed a bimodal distribution with one peak at 0.004% and another at 0.035% (Fig. 5b). The SNP rate of 14.9% of the genome regions was lower than 0.015%, and the SNP rate of the remaining regions was in range 0.015–0.6%. According to the Wright–Fish population genetic model, the effective population size is inversely proportional to the time to the most recent common ancestor (TMRCA), and TMRCA directly determines the heterozygosity, so this SNP distribution might indicate changes or bottlenecks of the effective population size in the panda population history.

---

**Figure 4 | Structure of the umami receptor TIR1 gene.** Two frameshift mutations occurred in the third and sixth exons (red) of the panda TIR1 gene. The third exon contained a 2-bp (‘GG’) insertion; the sixth exon contained a 4-bp (‘GTGT’) deletion.
and scaffolds of the panda genome were sufficiently long to allow us to carry out our gene prediction and comparative analyses, and our evaluation of the sequence quality showed that the panda genome sequence was of similar utility to sequences generated using traditional capillary-based Sanger sequencing. It should be noted that fewer active transposons and recent segmental duplications in the panda genome may have made it relatively easier to assemble than some other complex animal or plant genomes. Thus, the relative effect of these genomic features on de novo assembly from short reads will need to be assessed in the future in genomes that contain such features.

The panda genome is, to our knowledge, the first genome from the Ursidae lineage, and the second carnivore to be sequenced. Our comparative analysis showed that the panda had high genomic synteny with dog and human, but that there has been considerable divergence in the repetitive regions, most of which seems to result from recent transposable-element activity. Our comparison of the panda and other available mammalian genomes showed that most genes were conserved among all of these mammals, although we also identified several gene families that had expanded or contracted within each lineage. Of interest, our analysis of genes potentially involved in the evolution of the panda’s reliance on bamboo in its diet showed that the panda seems to have maintained the genetic requirements for being purely carnivorous even though its diet is primarily herbivorous. Furthermore, given our finding that some of the genes necessary for complete digestion of bamboo are missing from its genome, investigation of panda’s gut microbiome may be important for understanding its unusual dietary restrictions.

As well as providing a reference genome for use in future studies, we were also able to detect the most extensive heterozygous SNPs, indels and structural variation in the diploid genome, which will be important for ecological and population studies designed to address issues of panda survival.

METHODS SUMMARY

We constructed the paired-end DNA libraries with insert sizes larger than 2 kb by self-fragmentation of the DNA fragments and merging the two ends of the DNA fragment. We randomly fragmented the circularized DNA and enriched the fragments crossing the merged boundaries using magnetic beads with biotin and streptavidin. The sequencing process followed the manufacturer’s instructions (Illumina), and the fluorescent images were processed to sequences using the Illumina data processing pipeline (v1.1).

The genome sequence was assembled with short reads using SOAPdenovo software (http://soap.genomics.org.cn), which adopts de Bruijn graph data structure to construct contigs. The reads were then realigned to the contig sequence, and the paired-end relationship between the reads was transferred to linkage between contigs. We constructed scaffolds starting with short paired-ends and then iterated the scaffolding process, step by step, using longer insert size paired-ends. To fill the intra-scaffold gaps, we used the paired-end information to retrieve read pairs that had one read well-aligned on the contigs and another read located in the gap region, then did a local assembly for the collected reads.

Known transposable elements were identified using RepeatMasker (version 3.2.6) against the Repbase transposable element library (version 2008-08-01), and highly diverged transposable elements were identified with RepeatProteinMask by aligning the genome sequence to the curated transposable-element-related proteins. A de novo panda repeat library was constructed using RepeatModeler. Using evidence-based gene prediction, the curated and dog genes (Ensemble release 52) were projected onto the panda genome, and the gene loci were defined by using both sequence similarity and whole-genome synteny information. De novo gene prediction was performed using Genscan and Augustus. A reference gene set was created by merging all of the gene sets. The sequencing reads were mapped on the panda genome sequence using SOAPAligner, and heterozygous SNPs were identified by SOAPmap.

Received 19 August; accepted 24 November 2009.
Published online 13 December 2009.


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We are indebted to the faculty and staff of Beijing Genomics Institute at Shenzhen, whose names were not included in the author list, but who contributed to this team work. This project is funded by the Shenzhen Municipal Government and the Yantian District local government of Shenzhen. This project is supported by the National Natural Science Foundation of China (30725008). Oie Romer grant from the Danish Natural Science Research Council, the Solexa project (272-07-0196), and Danish Strategic Research Council (2106-07-0021). We thank C. Wang, X. Huang and L. Luo for their assistance in collecting the blood sample. We thank D. D. Pollock for repeat content estimation. We thank M. Shi, Y. Wang and M. Ting Wang for analysis of the replication related genes. We also thank L. Goodman for editing the manuscript.


**Author Information** This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession ACTA00000000. The version described in this paper is the first version ACTA00000000. The reads data have been deposited in the EBI/NCBI Short Read Archive under the accession SRA009203. All panda data have been released for public use and can be freely accessed at BGI official website: http://panda.genomics.org.cn/download.jsp. The entire dataset includes all usable reads, genome assembly, annotation of genes, repeat and non-coding RNA, as well as heterozygous SNP, indel and structural variation information. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Ju.W. (wangjian@genomics.org.cn) or Ji.W. (wangjian@genomics.org.cn).