Supplementary methods

Genetic background of panda Jingjing

We sequenced a 3-year old female panda named Jingjing, who came from the Chengdu Research Base of Giant Panda Breeding and was chosen as the model of 2008 Olympic Mascot. The captive breeding of pandas follows principles meant to maintain genetic polymorphism, so it is difficult to find a homozygous panda because mating of close relatives is not allowed. The genetic background of Jingjing's father comes solely from pandas in the Liangshan Mountain, and Jingjing's mother's genetic background has half from pandas in the Liangshan Mountain and half from the Minshan Mountain. Liangshan and Minshan are the two major locations of wild giant pandas. So Jingjing is a relatively good candidate for sequencing.

DNA library construction and sequencing

100 ml peripheral venous blood was phlebotomized from giant panda Jingjing and genomic DNA was extracted using Puregene Tissue Core Kit A (Qiagen). For short-insert (150 bp and 500 bp) DNA libraries, i.e. standard DNA libraries, we used the same manufacture's protocol (Illumina). Briefly, 5 µg of genomic DNA was fragmented by nebulization with compressed nitrogen gas. We then polished the DNA ends and added an "A" base to the ends of the DNA fragments. Next, the DNA adaptors (Illumina) with a single "T" base overhang at the 3' end were ligated to the above products. We then purified the ligation products on a 2% agarose gel, and excised and purified gel slices for each insert size (Qiagen Gel Extraction Kit). We required the insert size of a library to fall in a narrow range (peak – peak * 10%, peak + peak * 10%), in order to facilitate the assembly process.

For long (>= 2 Kb) mate-paired libraries, we used the manufacture's mate pair library kit (Illumina), 10-30 µg genomic DNA was fragmented by nebulization with compressed nitrogen gas, and then we used biotin labeled dNTPs for polishing, and gel selection for the main bands among 2 Kb, 5 Kb, and 10 Kb. The requirement of insert size range was similar to that used for the short insert (150 bp and 500 bp) libraries. For self ligation the DNA fragments were circularized, so the two ends of the DNA fragment were merged together, the linear DNA fragments were digested by DNA Exonuclease, and then the circularized DNA was fragmented again, followed by enrichment of the "merged ends" with magnetic beads using a biotin/streptavidin, then the ends were polished and "A" base and adaptors added. The sequencing was the same as for standard genomic DNA libraries.

For performance PE (paired-end) sequencing runs, we followed the manufacture's user guide (Illumina); cluster generation was performed using the Illumina cluster station; and the workflow was as follows: template hybridization, isothermal amplification, linearization, blocking, sequencing primer hybridization, and sequencing on the sequencer for Read 1. After the first read was completed, we prepared the second read as follows: denaturation, de-protection, re-synthesis, linearization, blocking, primer hybridization, and sequencing in the opposite direction of the dsDNA fragments. The average raw cluster density was about 100,000 clusters per tile, spanning 50,000 to 150,000 clusters per tile. Given a lane has 100 tiles, there were about 10 million raw clusters per a lane. The actual production was approximately proportional to the raw cluster density in this range; however, it decreased if it had a higher cluster density, and the sequencing errors also increased significantly.

Quality checking on the library and read filtering

Each library was first sequenced in one or two lanes for quality checking, and then a decision was made as to whether it qualified for large-scale sequencing. In total, we built 37 DNA libraries to decrease the risk of non-randomness. If a library had a problem of abnormal base bias due to bad enzyme or primer degradation, we took it as a non-qualified library and did not assign it for any further sequencing. There were 31 qualified libraries, and we arranged relatively more lanes with libraries of higher quality.

The raw reads generated from the Solexa-Pipeline included some artificial reads that were caused by base-calling duplicates and adapter contamination. We filtered these artificial reads to get a clean usable reads set. The definition and method were as follows: (1) Base-calling duplicate, this is a unique characteristic for each lane, caused by the solexa-pipeline, and they are not real sequences. The higher the raw cluster density, the more severe this problem is. The redundant reads were filtered at a threshold of euclid distance <= 3 and a mismatch rate of <= 0.1. We observed that the average rate of base-calling duplicates for each lane was about 0.83%, ranging from 0.00% to 8.52%. (2) Adapter contamination, another unique characteristic of the specific library, is caused by DNA adaptor dimerization, the empty loading or too small an insert size (less than the read length). If Read 1 contained a 3'-adapter, then Read 2 should contain a 5'-adapter. The reads were filtered at a threshold if both Read 1 and Read 2 contained an adapter >= 10bp with a mismatch rate <= 0.1. We observed that the average rate of adapter contamination for each library was about 0.10%, ranging from 0.01% to 1.28%.

Extraction of high quality reads for assembly

The quality requirements for *de novo* sequencing is far higher than for re-sequencing, because sequencing errors can create difficulties for the short-read assembly algorithm. We therefore carried out a stringent filtering process. In addition to filtering base-calling duplicates and adapter contamination, we also included additional and more stringent filtering measures as follows: (1) 17 lanes were excluded totally, because of their overall low quality. (2) Reads from 42 lanes were trimmed at the 3-end to remove low-quality sequences. (3) For long insert-sizes (>=2 Kb) libraries, the duplicated reads that were generated by PCR amplification in the library construction process were filtered, because the duplication rate for these libraries is much higher than for short insert (150 bp~500 bp) libraries. By eliminating the duplicated reads, we ensured the high accuracy of scaffold construction with these large mate-paired reads. (4) We also checked the individual reads in all lanes, and filtered those reads with a significant excess of "N" and low-quality bases.

The algorithm of SOAPdenovo short-read assembly

To assemble large genomes with the massive short reads, SOAPdenovo[1] (http://soap.genomics.org.cn/soapdenovo.html) adopted the *de Bruijn* graph data structure, which was introduced in the EULER assembler[2].

Prior to assembly, the sequence errors were corrected based on *K*-mer frequency information. For the panda genome assembly, we chose *K*=17 bp, and corrected sequencing errors for the 17-mers with a frequency lower than 4. In summary, we corrected 8.4% of the reads and 0.2% of the bases. The total, the number of distinct 27-mers (we used 27-mer in graph construction and assembly) was reduced from 8.62 billion to 2.69 billion (3.2 times smaller) through this error correction step.

After error correction, we loaded the short-insert–size paired-end reads into RAM to construct a *de Bruijn* graph, storing the overlap information among the short reads. We then followed the steps of clipping tips, merging bubbles, removing low-coverage links, and resolving tiny repeats. For the panda genome assembly, we chose *K*=27 bp when constructing the *de Bruijn* graph. We removed 72 million tip nodes and merged 2.6 million bubbles during the graph simplification process. Finally, we obtained the contig sequences by conjoining the *K*-mers in an unambiguous path. The contigs were broken into fragments at the boundaries of repeat ambiguous connections. By reporting contigs with >= 100 bp, the N50 and N90 contig size was 1,483 and 224 bp, respectively.

We realigned all the reads onto the contig sequences and obtained 80% of all the aligned paired-end reads. We then calculated the amount of shared paired-end relationships between each pair of contigs, weighted the ratio of consistent and conflicting paired-ends, and then constructed the scaffolds step by step, from short insert-sized paired-ends, to long distant paired-ends. We required at least 3 consistent read pairs to form a connection. Using 150- and 500-bp insert-sized data, we obtained an N50 and N90 scaffold size of 33 and 8 Kb, respectively; by adding 2 Kb insert-sized data, we obtained an N50 and N90 scaffold size of 229 and 45 Kb; by adding 5-Kb inserted-sized data, N50 and N90 size scaffolds were improved to 582 and 127 Kb; finally, using 10-Kb insert-sized data, we obtained scaffolds N50 1,282

Kb and N90 313 Kb. In principle, the scaffold size could have been further improved by using even more distant insert-sized paired-end data, such as fosmid ends (~35 Kb) and BAC ends (100~150 Kb).

To close the gaps inside the constructed scaffolds, which were mainly composed of repeats that were masked during scaffold construction, we used the paired-end information to retrieve the 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 these collected reads. We constructed a De Bruijn graph (K=27) with the reads in the gaps and the contig ends on both sides in a manner similar to the contig construction process. If an unambiguous path was found between those two contig ends, we filled the gap with the path sequence. For tandem repeats, it is difficult to know the exact number of repeat units based on 27-mers, so we used the read sequences to improve gap closure of these tandem repeat regions. Our longest reads were 75 bp, with about 20X sequence coverage. We checked each read in the gaps to find one that had unambiguously mapped ends (> 10 bp) on both sides of the contigs, and then filled the gap with the read sequence. Most of the small tandem-repeat gaps were correctly filled by this way. For gaps containing longer tandem repeats that could not be resolved by read sequences, we filled them with two repeat units together with a string of “N”s, to indicate that there is a tandem repeat with an unknown number of units. We closed 97.2% of the intra-scaffold gaps, or 80.5% of the sum gap length. The contig N50 size grew from 1,483 bp to 39.9 Kb and the genome coverage was improved from 84.2% to 93.6%.

Repeat annotation

1) Identification of known TEs

We first identified known TEs using RepeatMasker (version 3.2.6)[3] against the Repbase[4] TE library (version 2008-08-01), and then executed RepeatProteinMask[3], a new software program in the RepeatMasker package, which identifies TEs by aligning the genome sequence to a self-taken curated TE protein database.

2) *De novo* repeat prediction

We constructed a *de novo* panda repeat library using RepeatModeler[3], at the heart of which are two complementary programs RECON[5] and RepeatScout[6], and used the default parameters. The generated results were consensus sequences and classification information for each repeat family. Then RepeatMasker was run on the genome sequences, using the RepeatModeler consensus sequence as the library.

3) Tandem repeats

We identified the non-interspersed repeat sequences using RepeatMasker with the "-noint" option, including Simple_repeat, Satellites, and Low_complexity repeats. We also predicted tandem repeats using TRF[7], with parameters set to "Match=2, Mismatch=7, Delta=7, PM=80, PI=10, Minscore=50, and MaxPeriod=12".

Gene annotation

1) Homology based gene prediction

We have built a pipeline to project the human and dog genes (Ensembl release 52) onto panda genome, which included 6 steps:

(a) Rough alignment. We aligned the protein sequences of the human and dog (the longest translations were chosen to represent each gene) to the panda genome by TblastN at E-value 1e-5, and grouped all the HSPs into gene-like structures by genBlastA[8].

(b) Precise alignment. We first cut out the target gene fragments in the genome by extending 500 bp at both ends of the alignment regions, included the intron regions, then aligned the parent protein sequences to these DNA fragments by Genewise[9].

(c) Transcript clustering. We clustered all the predicted transcript structures by genomic overlap with a cutoff of more than 50 bp. For each gene locus, the transcript supported by the whole genome synteny (Blastz/chain/net) was preferred, otherwise the transcript with the best aligning rate to its parent protein was chosen.

(d) Building gene-scaffold. If one gene mapped to more than one scaffold, we tried to build a gene-scaffold to complete the gene structure, by referring to the whole genome synteny. The gene segments distributed on two or more neighboring syntenic scaffolds were conjoined in order.

(e) Filtering pseudogenes. There are two types of frame errors, frame shift and inner stop codons, that mark pseudogenes. We filtered the single-exon genes that were derived from retro-transposition and contained a frame error. For multi-exon genes that were not supported by whole genome synteny, >=3 frame errors was required; while for those supported, >=8 frame errors was required.

(f) UTR attachment. We mapped the human mRNAs to the panda genome using Blat and selected the alignment with the longest matching length for each mRNA. We filtered alignments with identity less than 80% or aligning rate less than 50%. We then compared initial/terminal exons of a panda gene to each exon of the overlapping mRNA alignment to find the 5'/3'-UTR.

2) *De novo* gene prediction

We used two *de novo* prediction software programs: Genscan[10] and Augustus[11], with gene model parameters trained from *Homo sapiens*, and filtered partial genes and small genes that had less than 150 bp coding length. Then we aligned the predictions to a TE protein database using BlastP with E-value 1e-5 and filtered TE-derived genes that had more than 50% aligning rate.

3) Reference gene set

The homology-based and *de novo* gene sets were merged to form a comprehensive and non-redundant reference gene set. We clustered the genes from all the input sets with a cutoff of genomic overlap greater than 50 bp for each gene locus; the human-derived gene was preferred, then, if no human gene mapped, the dog-derived genes were used, finally, if no homologous gene mapped, the *de novo* prediction was used. We had a much stricter cutoff for *de novo* genes than for homology genes. The one with the larger CDS from Genscan and Augustus was chosen and was required to have more than 30% aligning rate to the SwissProt/TrEMBL database and had to have more than 3 exons.

4) Gene Function annotation

For the panda reference genes, we annotated the motifs and domains by InterPro[12] against publicly available databases including Pfam, PRINTS, PROSITE, ProDom, SMART, and PANTHER. The description of the gene products were presented by Gene Ontology[13], which was retrieved from the results of InterPro. We also mapped the panda reference genes to KEGG[14] pathway maps by searching the KEGG databases and finding the best hit for each gene. We then corresponded them to the pathway map and highlighted them with a distinguishable color to assist the pathway analysis.

ncRNA annotation

1) Identification of tRNA genes

The tRNA genes were predicted by tRNAscan-SE[14] with eukaryote parameters. If more than 80% length of a tRNA gene was covered by the SINE TEs, then it was defined as SINE masked. The tRNA identity to human was calculated on Muscle[15] global alignment.

2) Identification of rRNA genes

The rRNA fragments were identified by aligning the rRNA template sequences from the human genome using BlastN at E-value 1e-5, with cutoff of identity >= 85% and match length >= 50bp.

3) Identification of other ncRNA genes

The miRNA and snRNA genes were predicted by INFERNAL[15] software against the Rfam[16] database (release 9.1, 1372 families) with Rfam's family-specific "gathering" cutoff. To accelerate the speed, we performed a rough filtering prior to INFERNAL, by BlastN against the Rfam sequence database under E-value 1. The miRNA predictions were first aligned against the mature sequences of human and dog from miRBase[17] (release 13), allowing one base mismatch, and then aligned against the precursor sequences, requiring more than 85% overall identity. The snoRNA predictions were aligned to human H/ACA and C/D box snoRNAs and Cajal body-specific scaRNAs from snoRNABase[18] (version 3), and required a cutoff of 85% overall identity. The spliceosomal RNA predictions were aligned to the Rfam sequence database, and required a cutoff of 90% overall identity.

Pairwise whole genome alignment

Pairwise whole genome alignment among panda, dog and human was carried out using blastz[19], with the parameters: C=2, T=2, H=2000, Y=3400, L=6000, and

K=2200. Then the Chain/Net package was used for post treatment. The panda genome was masked with RepeatMasker repeats at “-s” setting and TRF tandem repeats of period <= 12. The dog (CanFam2.0) and human (hg18) repeat-masked genomes were downloaded from UCSC (http://genome.ucsc.edu).

Multiple whole genomes alignment

The 5-way whole genome multiple alignment, that included human (hg18), dog (CanFam2.0), panda, mouse (mm9), and rat (rn4), was generated using multiz[20] following the topology of species tree. The human genome was set to be the reference, and for input pairwise alignments, we carried out in-house generation of the human versus dog and human versus panda alignments, while the human versus mouse and human versus rat alignments were download from UCSC (http://genome.ucsc.edu).

Detection of conserved non-coding regions

Phastcons[21] was adopted to identify conserved elements with conservation scores, given a multiple alignment and a phylo-HMM. A phylo-HMM consisting of two states was assumed: a "conserved" state and a "non-conserved" state. The parameter settings were “--target-coverage 0.3 --expected-length 45 --rho 0.31”, the phylogenetic model for non-conserved regions was produced by phyloFit in the PHAST package.

Detection of chromosome breakpoints between panda and dog

We generated a whole genome pair-wise alignment between panda and dog using the blastz/chain/net[19] software on the repeat-masked genomes and identified clusters of unique alignments with well-defined order and orientation. These clusters were defined as syntenic segments. The syntenic segments retained the primary orientation of the alignments on which they were based. The intra-chromosomal breakpoints were primarily caused by orientation inversion of syntenic segments, while the inter-chromosomal breakpoints possibly indicated the occurrence of recombination of different chromosomal fragments during evolutionary history. As the panda assembly was generally fragmental, inter-chromosomal breakpoints were detected only on the panda scaffolds that each mapped primarily to two dog chromosomes. We tested different cutoffs (5 Kb to 100 Kb) for minimum syntenic segment size and observed various counts of intra-chromosomal and inter-chromosomal breakpoints. We then performed the panda vs human and dog vs human whole-genome pair-wise alignments using the same method as for panda vs dog, and used the human genome as an outgroup to differentiate whether panda or dog changed with regard to the rearranged regions.

Detection of recent segmental duplication in the panda genome

We analyzed the panda genome assembly using two complementary genome-wide approaches designed to detect genomic duplicates >1 kb length and >90% sequence identity, as follows:

1) Whole Genome Analysis Comparison (WGAC).

We applied Blastz[19], a BLAST-like implantation specifically designed for long genomic sequences alignments, instead of previous BLAST-based whole-genome assembly comparison (WGAC)[22], to identify all pairwise alignments >1 kb length and >90% identity within the panda assembly. Self-versus-self blastz alignment was performed using the repeat-masked genome sequence, and followed by chaining of well-ordered neighboring alignments. We first obtained the seeding segmental duplications (non-repeat alignment length >500bp and overall identity >85%), and then reintroduced the masked repeat regions to perform optimal global alignment to refine the alignment identity and define the boundaries of segmental duplications more accurately. The resulting alignments that extended to >1Kb length and had >90% sequence identity were deemed recent segmental duplications.

2) Whole Genome Shotgun Sequence Detection (WSSD).

As large and high-identity duplications are frequently collapsed within whole genome shotgun (WGA) short-read assemblies, they are difficult to detect by the WGAC method[22]. To resolve this issue, we performed a whole genome shotgun sequence detection (WSSD) by aligning all the Illumina GA reads onto the panda genome assembly using SOAPaligner[23]. Each read was mapped to its best genomic locus, for reads with repetitive hits, one of the best hits was randomly chosen. We allowed 3 mismatches for read lengths of 30~45 bp and 5 mismatches read length 50~75 bp, the identity allowed for mapping was primarily > 90%, which ensured that reads from recently duplicated segments could be mapped to the collapsed segments, and thus could be detected based on a significant excess of WGS read depth-of-coverage. We performed a two-step process: First, we scanned the whole genome according to the depth of each base, and connected the consecutive bases whose depth was >100 (between 65 and 130, the estimated whole genome sequencing depth is 65X) to form the seeding blocks. Second, we connected the neighboring seeding blocks to form the high-depth segments with >1 Kb length and average depth >100X, which were deemed as recent segmental duplications. Those segments that were primarily comprised of TEs (>90% in content) were filtered. Given the sequencing depth is affected by GC content, both very-low and very-high GC regions would have relatively low depth, so the duplication on these regions might remain undetected.

Construction of mammalian gene families with Treefam method

We used the Treefam's methodology[24] to define a gene family as a group of genes that descended from a single gene in the last common ancestor of considered species. A pipeline to cluster individual genes into gene families and perform phylogeny analysis was built, which utilized many of Treefam's software as well as some in-house programs. The pipeline includes 4 main steps:

1). Data preparation. For panda, the reference gene set was used. The protein-coding genes for 8 eutheria speices (*Canis familiaris*, *Felis catus*, *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, *Equus caballus*) and one outgroup species (*Monodelphis domestica*), were downloaded from Ensembl (http://www.ensembl.org) release 52, the longest translation was chosen to represent each gene, the cds, and protein sequences were made to be consistent, and genes shorter than 30 aa were filtered out.

2). Assignment of pairwise relation, i.e. graph building. BlastP was used on all the protein sequences against a database containing a protein dataset of all the species under E-value 1E-10, and conjoined fragmental alignments for each gene pairs by Solar. We assigned a connection (edge) between two nodes (genes) if more than 1/3 of the region aligned to both genes. An Hscore that ranged from 0 to 100 was used to weigh the similarity (edge). For two genes G1 and G2, the Hscore was defined as score(G1G2) / max(score(G1G1), score(G2G2)), the score here is the BLAST raw score.

3). Extracting gene families, i.e. clustering by Hcluster_sg. We used the average distance for the hierarchical clustering algorithm, requiring the minimum edge weight (Hscore) to be larger than 5, and the minimum edge density (total number of edges / theoretical number of edges) to be larger than 1/3. The clustering for a gene family would also stop if it already had one or more of the outgroup genes.

4). Phylogeny and orthology analyses. We performed multiple alignments of protein sequences for each gene family by Muscle[25], and converted the protein alignments to CDS alignments using a Perl script. We built phylogenetic trees using Treebest, which takes advantage of both codon-based and aa-based algorithms (nj-dn, nj-ds, nj-mm, phyml-aa and phyml-nt), and followed this by adjusting to the topology of species tree to form a more accurate consensus tree. We infered all the orthology (descended from speciation) and paralogy (descended from duplication) gene relations from the gene phylogeny tree.

Note that the software Solar, Hcluster_sg, and Treebest can be freely download from the sourceforge website (http://treesoft.svn.sourceforge.net/viewrc/treesoft/). The 1:1:1 single-copy gene families, which contain one gene of each species, were extracted and used for further phylogeny, molecular clock, and selection strength analyses. The gene families data were also used to facilitate the Panda-specific characteristics analysis.

Phylogeny reconstruction for 10 mammal species

In total, 7,034 single-copy gene families were defined using the Treefam methodology. Gene families, where the alignment lacked information site or had too many gaps (caused by potential incorrectly defined orthologs), were filtered for further analysis. The remaining single-copy gene families were used to reconstruct the phylogeny. 4-fold degenerate sites were extracted from each family and concatenated to one supergene for one species. Modeltest[26] was used to select the best substitution model (GTR+gamma+I) and Mrbayes[27] was used to reconstruct the phylogenetic tree. The chain length was set to 50,000,000 (1 sample/1000 generations) and the first 1,000 samples were burned in. Two independent runs were carried out and reached the same result. Branch-specific dN and dS were estimated with codeml in PAML with branch model[28].Transition/transversion rate ratio was estimated as a free parameter. Other parameters were set with the default setting.

Substitution rate in the giant panda lineage

20 sequenced nucleic genes of the American black bear were obtained from NCBI **(**http://www.ncbi.nlm.nih.gov**)**. Modeltest[26] was used to select the best substitution model (GTR+gamma+I), and Mrbayes[27] was used to reconstruct the phylogenetic tree. The chain length was set to 5,000,000 (1 sample/1000 generations) and the first 1,000 samples were burned in. Two independent runs were carried out and reached the same result. Bayesian molecular dating was adopted to estimate the neutral evolutionary rate and species divergence time using the program MULTIDIVTIME, which is implemented in the Thornian Time Traveller (T3) package (ftp://abacus.gene.ucl.ac.uk/pub/T3/). The calibration time (90 Mya) from human-dog divergence was achieved from the TimeTree database[29].

Phylogenetic and Evolutionary Analysis of FSHB gene

Orthologous and paralogous nucleotide coding sequences of FSHB from 11 mammalian species were downloaded from the Ensembl database. They were aligned with the two giant panda FSHB sequences (Giant Panda-FSHB1 and Giant

Panda-FSHB2) for building the phylogenetic tree; chicken FSHB sequences were used as the outgroup. Maximum likelihood method, GTR+Γ_4+G substitution model, and bootstrapping analysis of 1000 replications were used in the phylogenetic analysis as previously described. Branch-specific codon model[30], implemented in *PAML*[31], was used to estimate the selection pressure on each branch of the phylogeny, including the Giant Panda-FSHB2 branch.

Analysis of the pigmentation genes

Orthologous sets of genes from the dog, rat, mouse, macaque, chimp, and human were obtained from http://compgen.bscb.cornell.esu/projects/mammals-psg/[32]. Sequences were aligned using Clustal W2[33] and then adjusted by eye. Indels (insertions and deletions) and frameshift mutations were searched by eye in the panda sequences. Modeltest[34] 3.7 was implemented to select the model of sequence evolution. Phylogenetic trees were estimated with PAUP[35] 4.0 by constraining topologies to ((((human,chimp),macaque),(mouse,rat)),(dog,panda)) according to previous studies[36]. Positive selection in the panda was tested using PAML[37] 4.2. Two likelihood ratio tests (LRTs) were performed based on widely used branch-site models of codon evolution (Yang and Nielsen 2002). Positive selected sites were identified by the Bayes Empirical Bayes analysis[38].

Read mapping to the genome

All the usable reads were mapped to the scaffold sequences by SOAPaligner[23](version 2.18). The un-gapped alignment was performed first, with 3 mismatches allowed for read length <=45bp, and 5 mismatches allowed for longer reads. For those unmapped reads, a gap-tolerated alignment was performed, allowing one maximum continuous 6-bp gap on a read. For reads with repetitive hits, only one random best hit was reported. The mapped reads were divided into paired reads if their coordinate distances fell within a range (Library insert size ± 3*SD), otherwise they were put into single reads.

Heterozygous SNP calling

Heterozygous SNP were called by SOAPsnp[39] with the un-gapped aligned reads obtained from SOAPaligner alignments. A statistical model based on Bayesian theory and Illumia quality system was used to calculate a probability for each possible genotype at each position on the reference genome. At each position, the genotype was the allele type that had the highest probability and a rank sum test was applied to adjust the probability of heterozygote, and finally, a consensus sequence (CNS) was obtained. The final CNS probability was transformed to a quality score in the Phred scale.

The candidate SNP set was retrieved from the CNS, and 5 thresholds were used to post-filter unreliable SNPs: 1) requiring Q40 quality (quality score >= 40); 2) the overall sequencing depth be less than 130; 3) the approximate copy number of flanking sequences (< 2); 4) at least 5 uniquely mapped reads for each allele; and 5) the minimum distance that SNPs were away from each other (>= 5bp). The candidate length (denominator) for SNP rate calculation was also obtained by filtering the CNS file with a similar threshold.

Small Indel

The mapped reads that satisfied the pair-end requirements and contaiened alignment gaps at one end were used to identify indels. As the maximum gap length allowed was 6 bp, this limited the indels that could be detected in our study to those that were 1-6 bp in length. The PCR-duplicated reads that had the same outer coordinates in mapping were merged prior to looking for indels. Gaps that were supported by at least 5 gapped paired-end reads were extracted. An indel was called if there was at least 5 ungapped reads that crossed a possible indel, the ratio of gapped / ungapped fell in a range of 1/3~3, and the total number of mapped reads (gapped and ungapped) was less than 130.

Structural variation

We defined a read pair as a diagnostic paired-end (PE), if the two ends of a read pair could both be aligned but could not meet the pair-end insert size and/or orientation requirement. We grouped abnormally mapped paired-end reads into diagnostic PE clusters by looking for high density regions of abnormally mapped paired-end reads along the genome. Common structural variations, like deletions, insertions, and inversions, were examined and summarized into alignment models. We checked the diagnostic clusters to fit models to detect all possible SVs. To avoid false positive predictions, PE clusters with <10 pairs were discarded, the length of predicted variation was required to be larger than 100 bp, and the ratio of diagnostic PEs / Normal PEs had to fall in the range of 1/3~3.

Supplementary figures

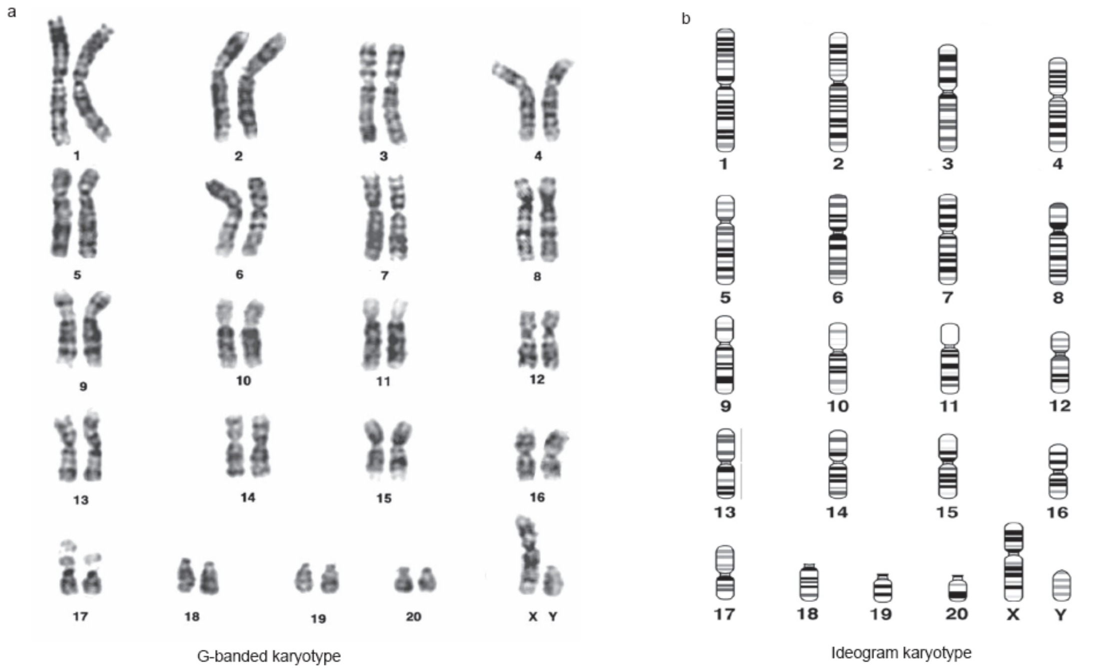
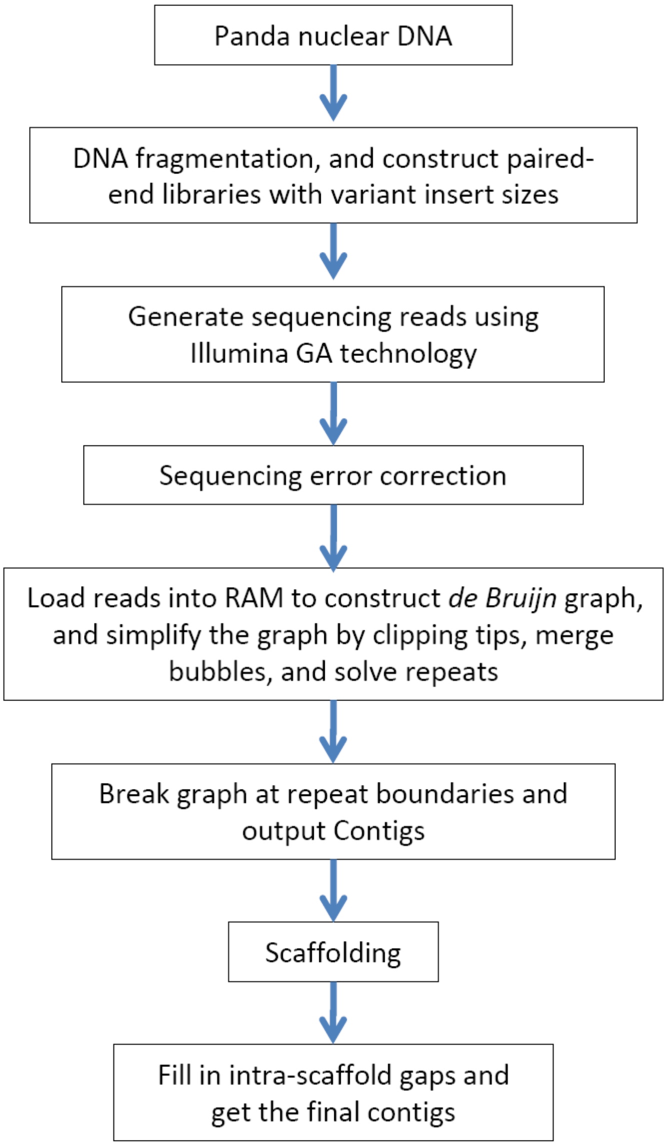


Figure S1 | a. G-banded and b. Ideogram karyotype for the panda genome (from Book “Atlas of mammalian chromosomes”, edited by Stephen J. O’Brien, Joan C. Menninger, and William G. Nash).



Panda nuclear DNA

DNA fragmentation, and construct paired-end libraries with variant insert sizes

Generate sequencing reads using Illumina GA technology

Sequencing error correction

Load reads into RAM to construct *de Bruijn* graph, and simplify the graph by clipping tips, merge bubbles, and solve repeats

Break graph at repeat boundaries and output Contigs

Scaffolding

Fill in intra-scaffold gaps and get the final contigs

Figure S2 | Flowchart of the panda genome *de novo* assembly.

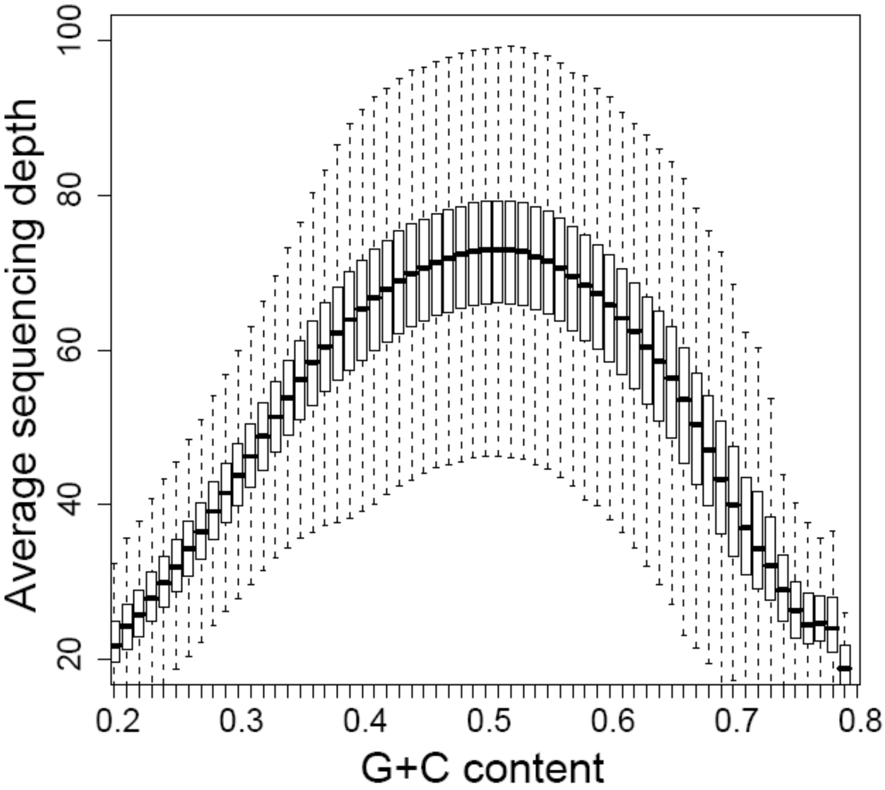


Figure S3 | Local GC content versus sequencing depth. We used 500-bp non-overlapping sliding windows along the assembled sequence to calculate GC content and average sequencing depth. The box plot was performed using the R package.

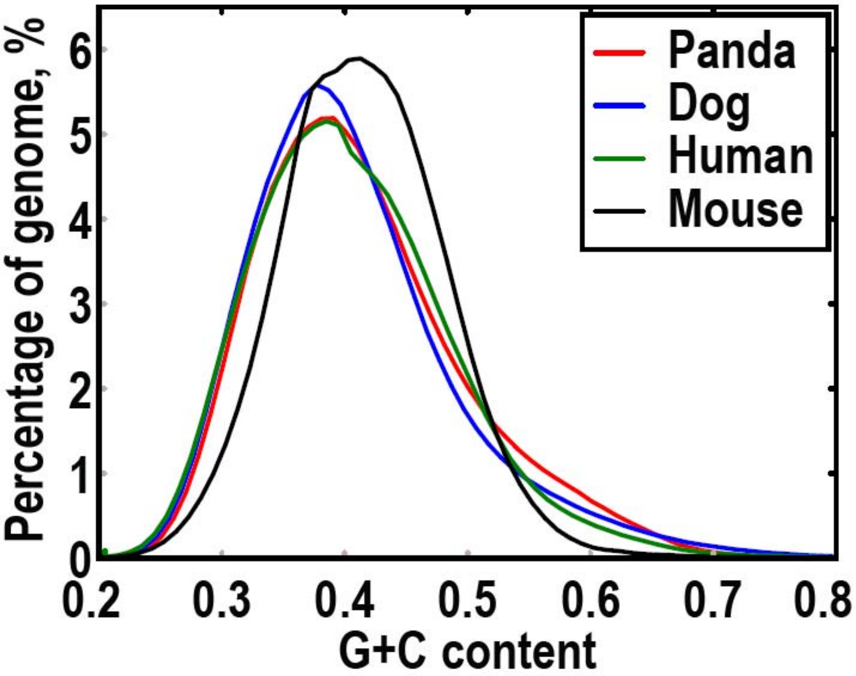
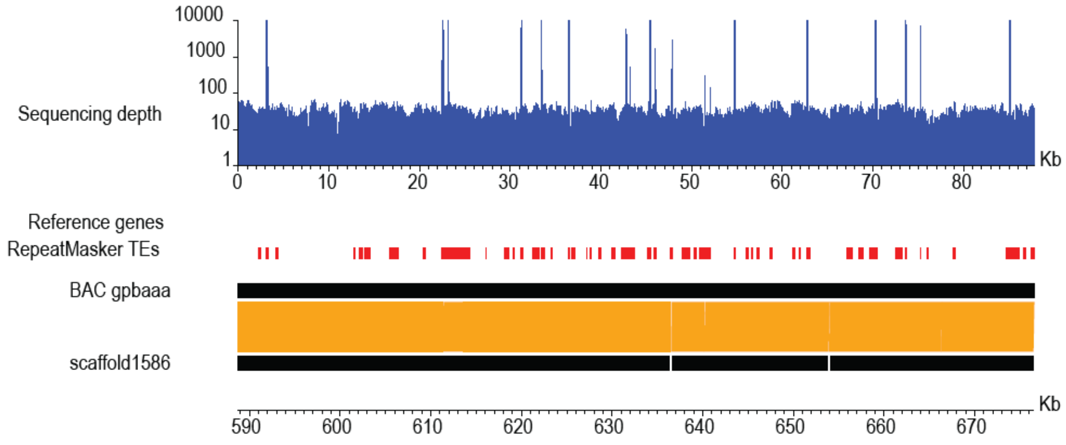
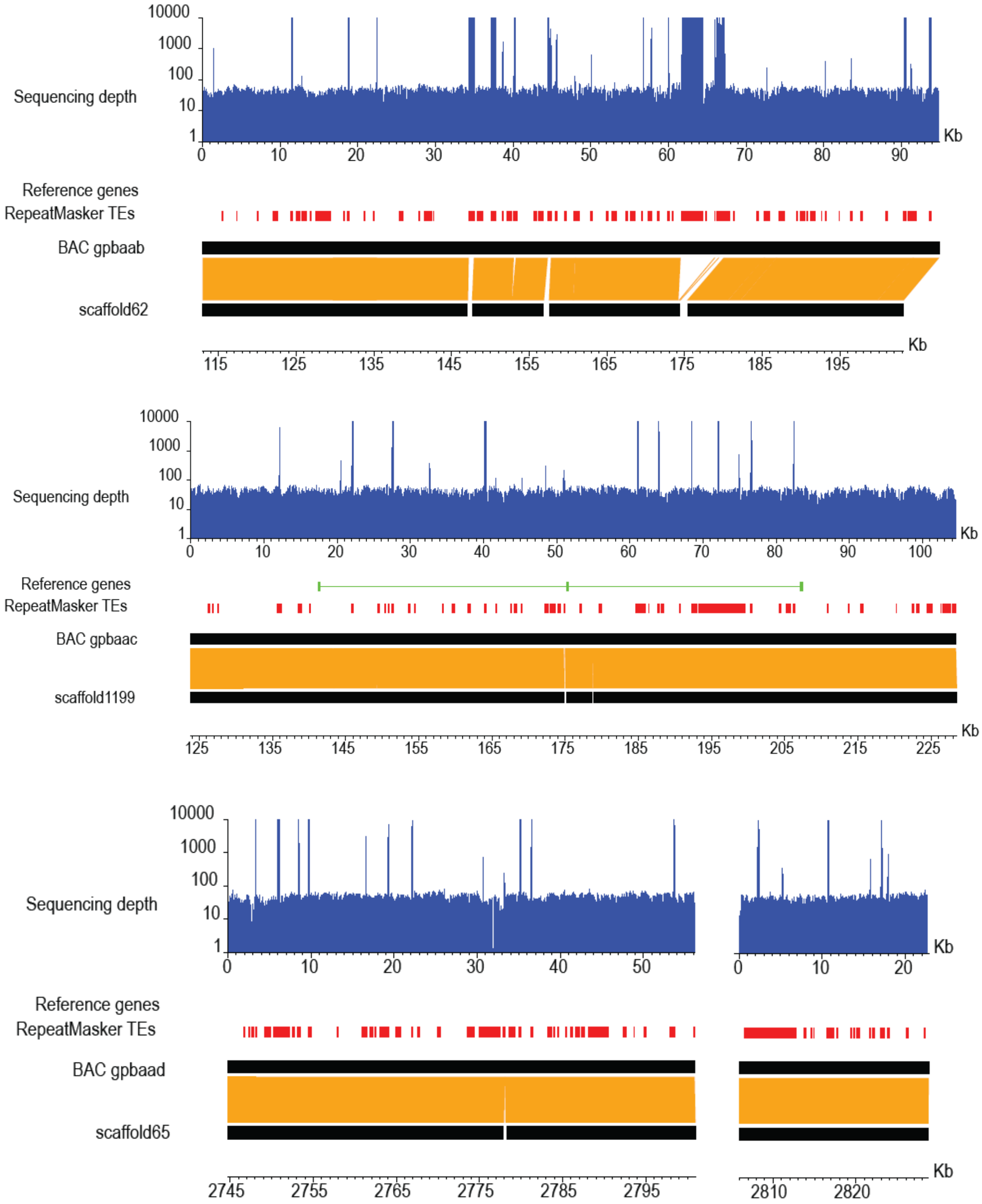


Figure S4 | Local GC content distribution of the human, mouse, dog, and panda genomes. We used 500-bp non-overlapping sliding windows along the genome.




Sequencing depth
10000
1000
100
10
1
0 10 20 30 40 50 60 70 80 90 Kb
Reference genes
RepeatMasker TEs
BAC gpbaab
scaffold62
115 125 135 145 155 165 175 185 195 Kb
Sequencing depth
10000
1000
100
10
1
0 10 20 30 40 50 60 70 80 90 100 Kb
Reference genes
RepeatMasker TEs
BAC gpbaac
scaffold1199
125 135 145 155 165 175 185 195 205 215 225 Kb
Sequencing depth
10000
1000
100
10
1
0 10 20 30 40 50
0 10 20 Kb
Reference genes
RepeatMasker TEs
BAC gpbaad
scaffold65
2745 2755 2765 2775 2785 2795
2810 2820 Kb

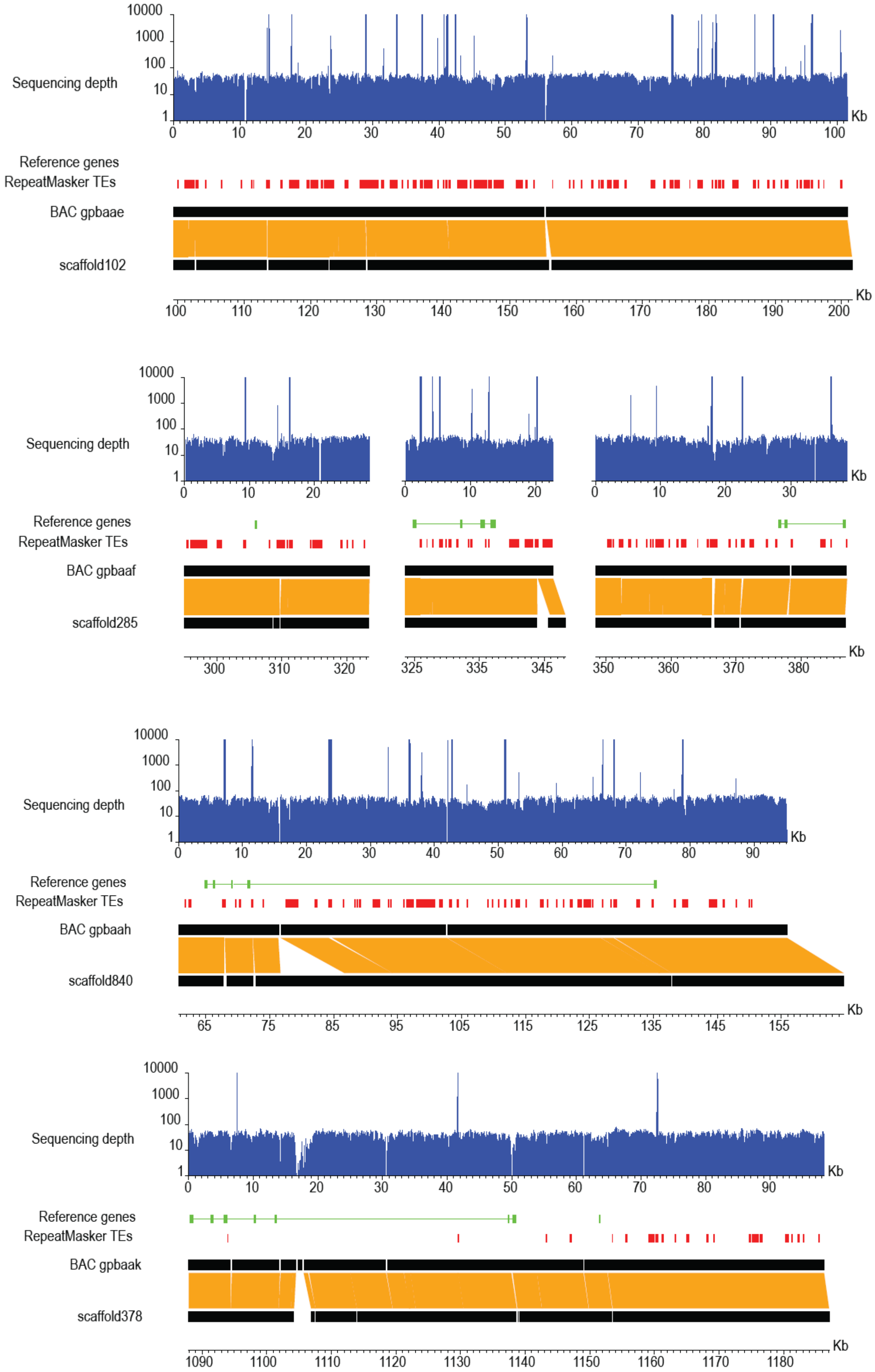


Figure S5 | Comparison of the assembled genome with 8 BAC sequences. The figure style for each BAC is same as for Figure 1b. Sequencing depth on the BAC was calculated by mapping the Illumina GA short reads onto the BAC sequence, here we performed single-end mapping and report all the repetitive hits. The predicted genes and annotated TEs on the BAC sequence are shown in green and red, respectively. The remaining unclosed gaps on the scaffolds and BACs are marked as white blocks.

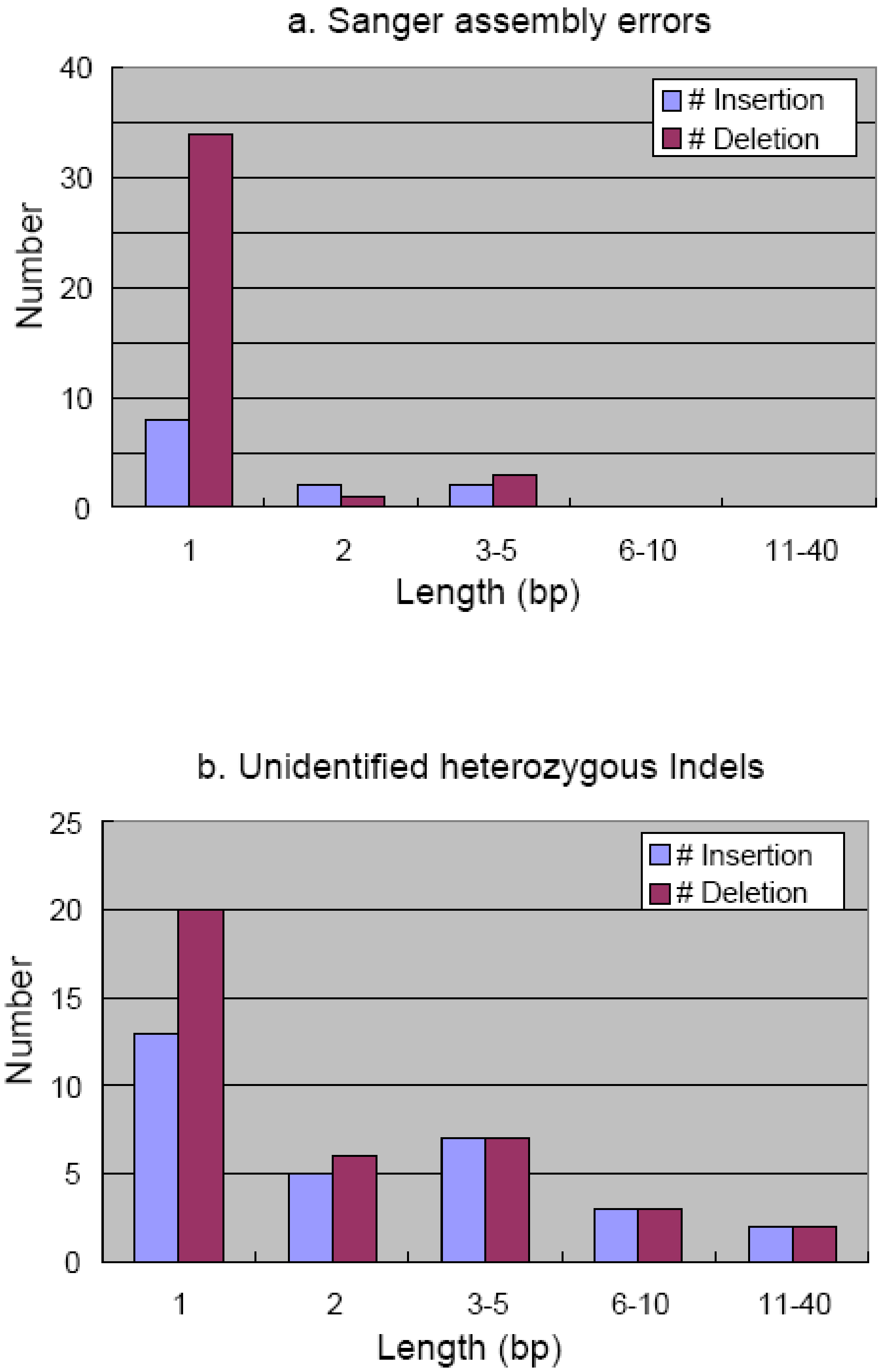
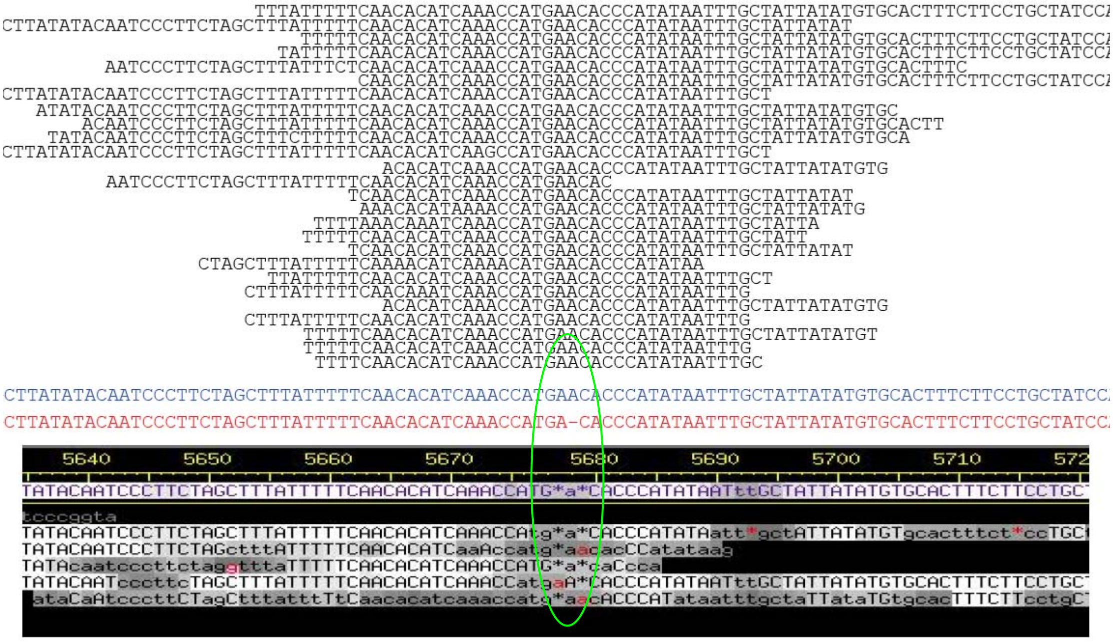


Figure S6 | Length distribution of insertion/deletion differences in comparison with BACs and scaffolds. We manually divided all the insertion/deletion differences into 2 categories: **a**. Sanger assembly errors and **b**. unidentified heterozygous Indels. The number of insertion and deletion differences is shown separately

a. Sanger assembly error



b. Unidentified heterozygous Indel

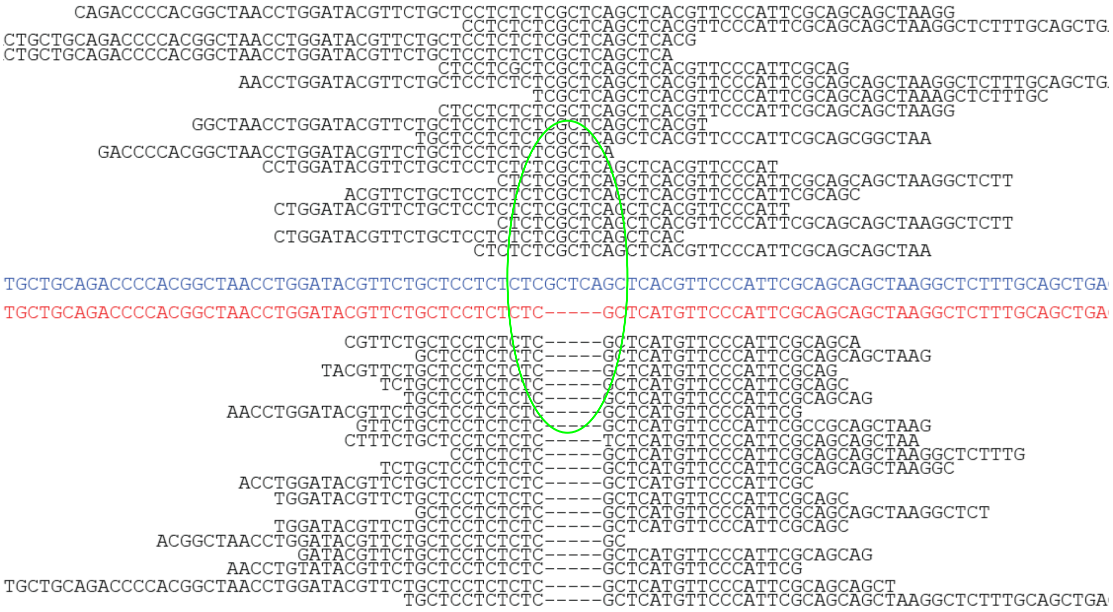


Figure S7 | Insertion/deletion differences in comparison of BAC and scaffolds.

a. An example of a Sanger assembly error. In the middle, there is a scaffold (scaffold1586, blue) versus BAC (gpbaaa, red) alignment shown, the coordinates of the different regions on scaffold1586 was 594,204, and on gpbaaa was 5,440. At the top of the figure, there are reads mapped (gap un-tolerated and pair-end mapping) to the scaffold. At the bottom, there is read assembly information from Consed software. A green circle highlights the difference in the region. Because there were many reads that were correctly mapped and that crossed different regions on the scaffold, but there were no reads could be mapped to the BAC, and the Consed panel showed an ambiguous number of bases "A" on the corresponding region, we inferred that there was a 1-bp deletion error on the BAC sequence.

b. An example of an unidentified heterozygous Indels. The middle of the figure shows a scaffold (scaffold378, blue) versus BAC (gpbaak, red) alignment, the coordinates of the different regions on scaffold378 was 1,118,512-1,118,516, and on gpbaak was 29,572-29,572. The top and bottom of the figure shows the reads mapped (gap un-tolerated and pair-end mapping) to the scaffold and BAC, respectively. Because there were many reads that correctly mapped and crossed the different regions of the scaffold and BAC, we inferred that both the scaffold and BAC assembly were correct, and that this was a 5-bp heterozygous Indel in the diploid panda genome.

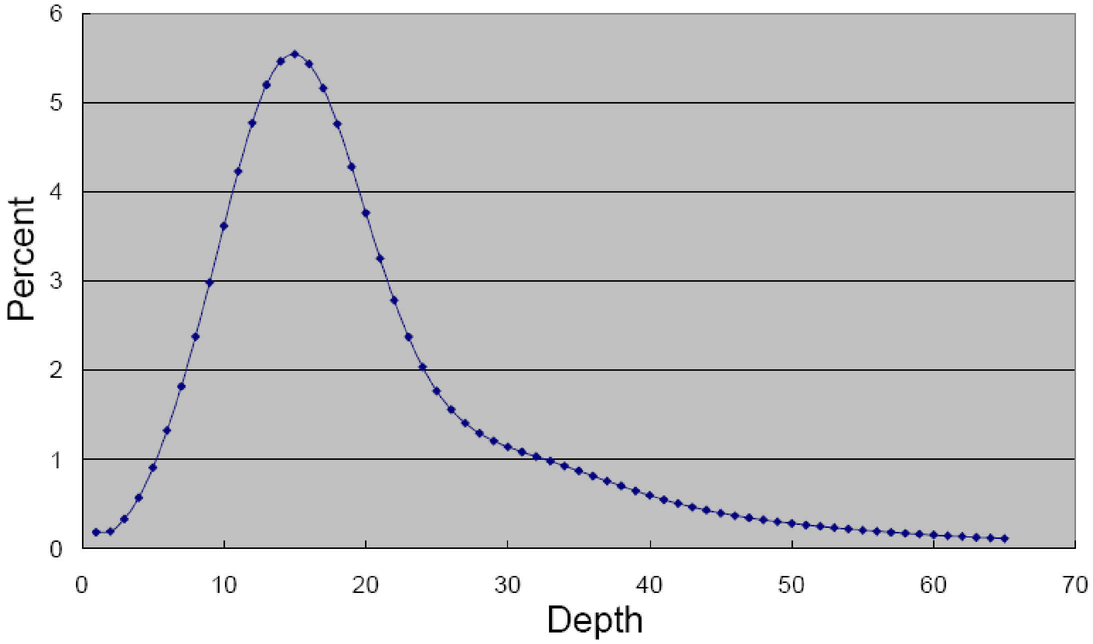


Figure S8 | Distribution of 17-mer frequency in the raw sequencing reads. We used all reads from the short insert-size libraries (<500bp). The peak depth is at 15X. The peak of 17-mer frequency (M) in reads is correlated with the real sequencing depth (N), read length (L), and kmer length (K), their relations can be expressed in a experienced formula: M = N * (L – K + 1) / L. Then, we divided the total sequence length by the real sequencing depth and obtained an estimated the genome size of 2.46 Gb.

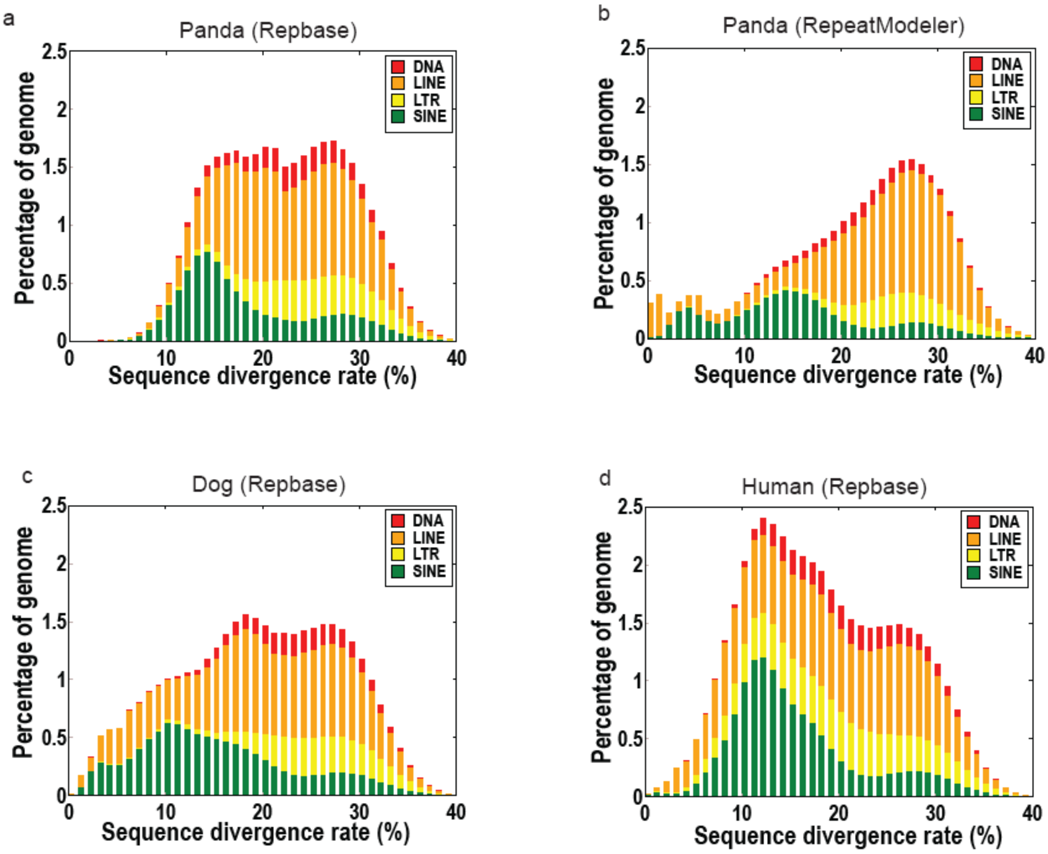


Figure S9 | Distribution of divergence rate of each type of TEs in the panda, dog, and human genome. The divergence rate was calculated between the identified TE elements in the genome and the consensus sequence in the TE library used (Repbase or RepeatModeler).

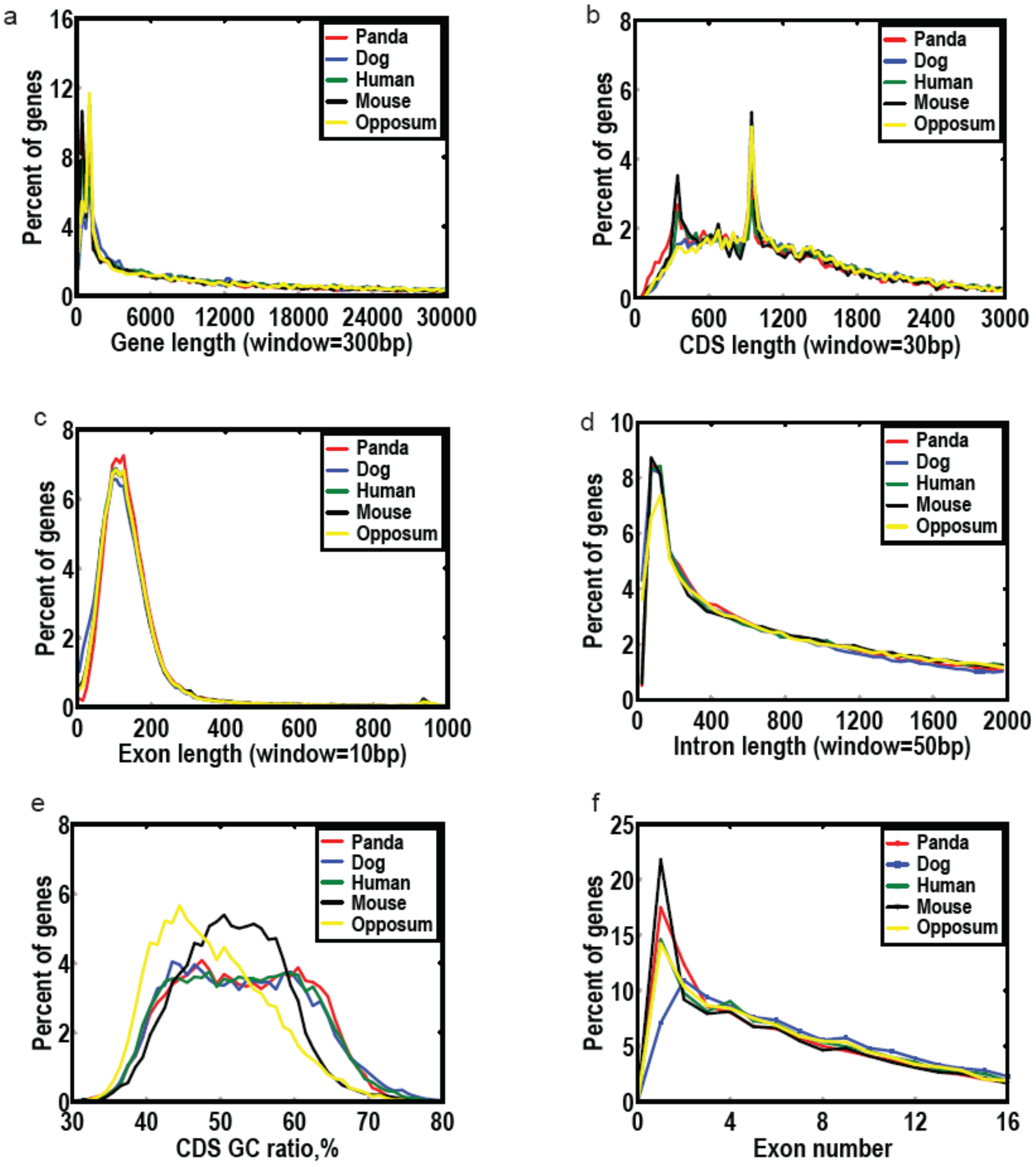


Figure S10 | Comparison of gene parameters among the sequenced mammalian genomes. No obvious unexpected differences were seen for panda, indicating the high quality of gene structure annotation.

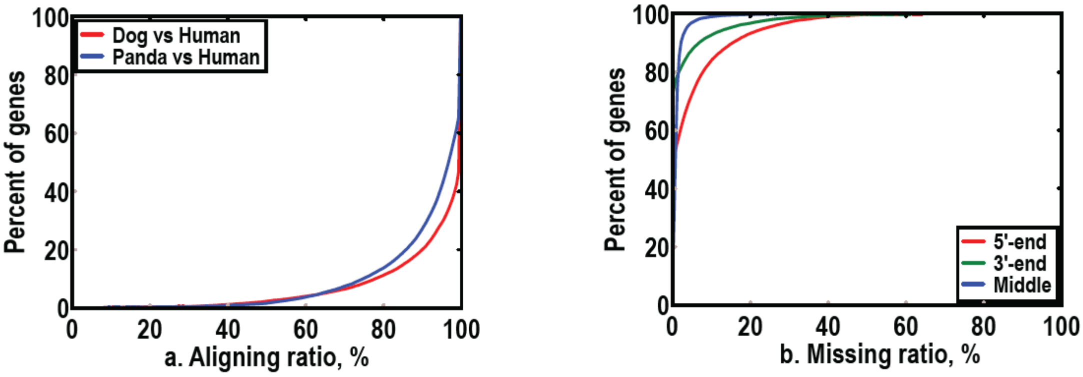


Figure S11 | Analysis of sequence completeness of the predicted genes. a. Alignment rate comparison between panda and dog using single-copy genes, both panda and dog genes were aligned to human genes, and the alignment rate was calculated for each pair of orthologous genes. **b.** The ratio of missing exons. The annotated panda genes were compared with the human genes, and the ratio of missing length was calculated on 5’-end, 3-end, and middle of genes.

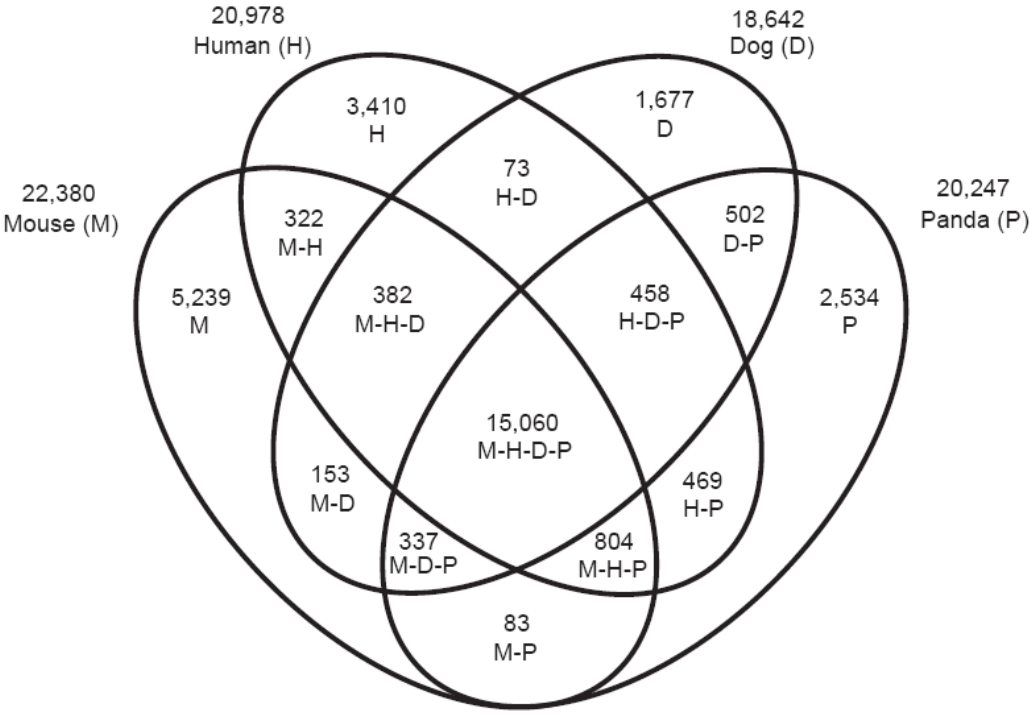


Figure S12 | Shared orthologous gene clusters among the panda, dog, human, and mouse genomes. Ensembl (v51) annotated genes were used for the human, mouse, and dog genomes. For genes with multiple alternative transcripts, the transcript with the best alignment was selected. Genes with lengths less than 100 bp were discarded. InParanoid was used to identify orthologous gene pairs, and then MultiParanoid was used to merge them into multiple species orthologous groups.

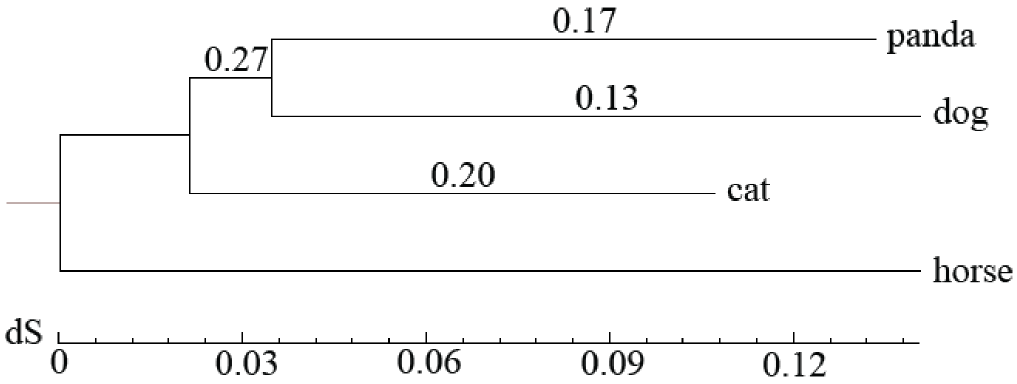


Figure S13 | Phylogenetic tree and dN/dS of T1R1 genes. The phylogeny tree was constructed with T1R1 gene families on synonymous sites, using the Neighbour-joining algorithm. Numbers on the branch represent dN/dS. It is 0.17 on panda lineage, which is larger than that of the dog lineage (0.13), indicating the recent death of the panda T1R1 gene.

a

```
Giant Panda-FSHB1    1  MKSVQFCFLFCCWRAICCKSCELTNITITVEKEECRFCISINTTWCAGYCYTRDLVYKDPARPNI  65
Giant Panda-FSHB2    1  ....RL.......K....S-............-..L......A...V.H...EN.....TG....  63
          Cat-FSHB   1  ..............V...........................A.....................N  65
         Dog-FSHB1   1  .R..R....L.....A.GAG.....V..A...........V.A................A...S.  65
         Cow-FSHB1   1  ..................R................G.....................R.......  65
         Cow-FSHB2   1  ...I..S...........S....N..N........S-.........V..TT......R.......  64
       Horse-FSHB1   1  .............K.V..N.........A......G.............................  65
       Horse-FSHB2   1  ....*V.*.S...K.V..N.....T.P.AM.....S....V.......H.LAL............  63
        Human-FSHB   1  ..TL..F......K....N.........AI.................................K.  65
   Chimpanzee-FSHB   1  ..TL..F......K....N.........AI..................H................  65
        Mouse-FSHB   1  ..LI.L.I..W.......H.........S...................................T  65
          Rat-FSHB   1  ...I.L.I.LW.L..V..H.........S.................E.................T  65

Giant Panda-FSHB1   66  QKICTFKELAYETVKVPGCAHQADSLYTYPVATECHCGKCDSDSTDCTVRGLGPSYCSFNEMKE  129
Giant Panda-FSHB2   64  ..T......T.DA.R....V..T...HM..A.....S....TN...YMM*....N....SK...  126
          Cat-FSHB  66  ..T......V.......................................Q.........S....  129
         Dog-FSHB1  66  ..T...R.......R.....RH....H...........R...............G....G.PQQ  129
         Cow-FSHB1  66  ..T......V...........H..............S......................RPLRQ  129
         Cow-FSHB2  65  ..T......V...........H.....................................R.I..  128
       Horse-FSHB1  66  ..T......V...........H...........A......N..................GD...  129
       Horse-FSHB2  64  ..T......V...........H....N......A......N..................GD...  127
        Human-FSHB  66  ..T......V....R......H...........Q.........................G....  129
   Chimpanzee-FSHB  66  ..T......V....R......H...........Q.........................G....  129
        Mouse-FSHB  66  ..V......V....RL....RHS....................................S....  129
          Rat-FSHB  66  ..V......V...IRL....RHS....................................G....  129
```

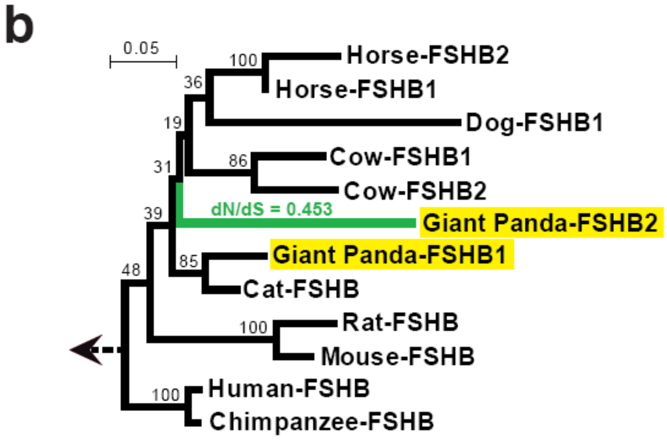


Figure S14 | FSHB genes and pseudogenes in mammals. **a**. Alignment of deduced amino acid sequences of the panda FSH beta subunit (FSHB1) gene with that of other mammalian FSHB (FSHB1) genes, or panda, cow, and horse putative pseudo-FSHB (FSHB2) gene. A putative pseudo-FSHB gene (FSHB2) was also found in the dog genome; however, the deduced partial amino acid sequence of the dog FSHB2 gene was too short (53 amino acids) and thus not included in this alignment. Dots indicate amino acids identical to the panda FSHB and dashes represent gaps in the sequence. Asterisks indicate premature stop codons in the coding region of pseudo-FSH beta subunit gene (FSHB2). **b**. Phylogenies of FSHB genes in different mammalian species. The tree was built using maximum likelihood method with GTR+Γ_4+G substitution model and bootstrapping analysis of 1000 replications applied (bootstrap values shown at the nodes). The two giant panda FSHB sequences are highlighted in yellow blocks. The lineage of Panda-FSHB2 is highlighted in green and shown with its estimated dN/dS. The dashed arrow shows the root by chicken FSHB. Note that the position of the Dog FSHB gene (Dog-FSHB1) is not consistent with the common species tree.

Supplementary tables

Table S1 | Statistics for each DNA library. Five categories of DNA libraries with various insert sizes were constructed. The low-quality libraries discarded in the *de novo* assembly are shadowed in the "Library ID" column.

Type	Library ID	Insert size (bp) §	Starting DNA amount (ug)	PCR Cycle #	Library concen-tration (ng/ul)	GA Lanes	GC%	Dupli-cate% *	Avg read length (bp) †	Usable reads (M) †	Usable bases (Mb) †	Avg read length (bp) ‡	High quality reads (M) ‡	High quality bases (Mb) ‡
150bp	PAfwDADAGAPE	112	5	12	8.7	6	41.4	7.7	44.0	68	2,994	44.0	61	2,688
	PAfwDADADAPE	113	5	12	9.4	1	40.5	6.6	44.0	7	288	35.0	6	195
	PAfwDADCAAPE	125	5	12	2.4	18	44.9	5.8	47.1	249	11,755	38.5	224	8,614
	PAfwDADBDAPE	134	5	12	13.0	6	43.4	5.8	44.0	72	3,152	41.7	65	2,695
	PAfwDADDCAPE	134	5	12	10.4	13	43.6	5.7	44.3	162	7,169	44.3	148	6,528
	PAfwDADCBAPE	144	5	12	5.3	22	43.7	5.6	45.9	323	14,825	40.7	293	11,931
	PAfwDADBGAPE	159	5	12	9.7	5	42.0	7.0	44.0	43	1,881	44.0	39	1,715
	PAfwDADBDBPE	160	5	12	11.0	5	40.4	5.6	44.0	68	3,010	40.3	61	2,455
	PAfwDADDAAPE	161	5	12	12.5	2	45.6	5.2	39.0	20	797	39.0	15	582
	PAfwDADBEAPE	173	5	12	12.9	3	43.2	5.4	44.0	40	1,755	44.0	36	1,592
	PAfwDADBFAPE	173	5	12	12.0	6	43.6	7.7	44.0	72	3,179	43.1	66	2,837
	PafwDAADBAAPE	201	5	12	9.5	1	45.6	4.4	40.0	14	574	40.0	13	529
	PAfwDADCEAPE	221	5	12	12.3	4	41.8	5.4	44.0	66	2,899	44.0	61	2,662
	PAfwDADCFAPE	221	5	12	11.8	6	41.8	4.8	44.0	102	4,468	38.3	84	3,235
	All 150bp libraries	159	5	12	10.1	98	43.3	6	45.0	1,306	58,745	41.2	1,171	48,258
500bp	PAfwDADFCAPE	380	5	12	10.1	8	43.1	5.7	66.6	122	8,142	71.1	83	5,890
	PafwDADJCAPE	403	5	12	11.5	9	45.3	6.0	66.4	120	7,980	67.0	82	5,466
	PAfwDADHCAPE	424	5	12	13.2	7	42.9	5.8	67.8	118	8,017	64.6	99	6,408
	PAfwDADHDAPE	442	5	12	6.1	7	44.0	6.4	72.6	120	8,700	72.0	98	7,047
	PAfwDADHAAPE	460	5	12	5.2	3	44.5	3.5	41.2	14	563	39.0	6	222
	PAfwDADKAAPE	487	5	12	13.2	1	44.9	6.8	39.0	17	668	39.0	16	620
	PafwDADICBPE	489	5	12	1.8	1	43.4	3.9	44.0	12	538	35.0	10	335
	PAfwDADICAPE	490	5	12	1.8	2	43.7	4.3	44.0	26	1,144	35.0	19	672
	PafwDADICCPE	492	5	12	6.5	4	43.3	6.3	68.8	76	5,191	56.1	65	3,654
	PafwDAADIAAPE	508	5	12	9.3	5	42.8	5.4	70.9	102	7,253	68.9	84	5,790
	PafwDAADIBAPE	528	5	12	12.7	5	42.8	5.5	68.5	107	7,326	67.9	85	5,793
	PafwDAADICAPE	561	5	12	12.8	5	45.3	4.9	67.0	83	5,576	59.9	38	2,290
	All 500bp libraries	472	5	12	8.7	57	43.7	5	66.6	917	61,099	64.6	684	44,187
2Kb	PAfwDAADWAAPE	1,768	10	18	1.5	6	44.8	16.7	72.2	85	6,129	52.2	45	2,339
	PAfwDAADWBAPE	2,059	10	18	7.1	6	44.8	13.5	72.4	109	7,903	57.8	100	5,752
	PAfwDAADWCAPE	2,383	10	18	12.8	5	44.7	15.0	71.6	92	6,561	75.0	73	5,508
	PAfwDADLAAPE	2,590	10	18	2.3	2	46.5	9.3	44.0	15	653	32.0	14	445
	PAfwDAADWDAPE	2,787	10	18	13.4	5	44.1	29.9	71.9	97	6,965	75.0	82	6,132
	All 2Kb libraries	2,317	10	18	7.4	24	44.7	17	71.0	397	28,211	64.4	314	20,177
5Kb	PAfwDADSAAPE	3,503	20	18	7.8	11	43.4	24.2	40.7	228	9,286	40.8	199	8,109

	PAfwDADLBAPE	3,760	20	18	1.6	1	46.2	8.4	44.0	12	536	32.0	12	373
	PAfwDADSBAPE	5,582	30	18	4.8	3	42.8	70.8	35.0	65	2,288	35.0	51	1,795
	PAfwDADTCAPE	7,412	30	18	4.3	7	42.9	53.1	35.0	199	6,974	35.0	153	5,358
	All 5Kb libraries	5,064	25	18	4.6	22	43.2	39	37.8	505	19,084	37.7	415	15,636
10Kb	PAfwDADTBAPE	9,293	30	18	6.5	7	43.8	62.7	35.0	159	5,576	35.0	125	4,364
	PAfwDADTAAPE	12,270	30	18	3.4	10	42.2	90.7	35.0	94	3,301	35.0	51	1,801
	All 10Kb libraries	10,782	30	18	5.0	17	43.2	77	35.0	254	8,878	35.0	176	6,166
All	All libraries	--	--	--	--	218	43.7	--	52.1	3,379	176,016	48.7	2,760	134,425

§ The range of paired-end insert sizes was estimated by mapping the reads onto the assembled genome sequence.
* To compare the duplicate level among different libraries, we defined and calculated the duplicate rate. We retrieved 1 Gb data from the first 40 bp for each library and calculated the read frequency. The reads with a frequency > 1 were called duplicated reads, and we defined the duplication rate as the count of duplicated reads / the count of total reads.
† Usable reads that were used in the estimation of the sequencing depth and identification of the heterozygous SNPs, Indels, and SVs. They were generated by filtering the base-calling duplicate and adapter contamination from the raw reads.
‡ High quality reads that were used for assembly. For all libraries, the low-quality reads were filtered. For the 2 Kb, 5 Kb and 10 Kb libraries, the duplicated reads generated by PCR in the library construction process were also excluded.

Table S2 | Summary of sequenced data of the panda genome. For our calculation of sequence coverage and physical coverage, we assumed a genome size of 2.4 Gb.

Paired-end libraries (bp)	Paired-end insert size (bp)	# libraries	# GA lanes	Avg reads length (bp) †	Sequence coverage (X) †	Physical coverage (X) †	Avg reads length (bp) ‡	Sequence coverage (X) ‡	Physical coverage (X) ‡
150	110~230	14	98	45	24.5	41	41	20.1	36
500	380~570	12	57	67	25.5	88	65	18.4	60
2K	1.7~2.8K	5	24	71	11.8	187	64	8.4	151
5K	3.7~7.5K	4	22	38	8.0	560	38	6.5	450
10K	9.2~12.3K	2	17	35	3.7	550	35	2.6	373
Total	All	37	218	52	73.3	1,426	49	56.0	1,070

† Usable reads that were used in the estimation of the sequencing depth and identification of the heterozygous SNP, Indel and SVs.
‡ High quality reads that were used for *de novo* assembly.
For the description of usable reads and high quality reads, refer to Table S1.

Table S3 | Statistics of tandem repeats in panda, dog and human genome. We identified tandem repeats (using TRF software) with period size less than 15 bp and length larger than 30 bp. Two sets of statistics were presented using two cutoffs:

Loose tandem repeats, requiring that percent of matches larger than 90% and percent of Indels less than 10%; Exact tandem repeats, requiring that percent of matches equal to 100% and percent of Indels equal to 0%.

	Loose tandem repeats			Exact tandem repeats		
Species	Number	Length (bp)	% in genome	Number	Length (bp)	% in genome
Panda	103,200	3,887,194	0.17	32,776	1,165,627	0.05
Dog	285,595	12,637,171	0.50	103,221	5,070,092	0.20
Human	201,521	8,047,062	0.26	88,471	3,395,372	0.11

Table S4 | Comparison of assembled scaffolds and 26 panda mRNA gene sequences in GenBank. The known panda mRNA gene sequences were downloaded from GenBank. The redundant items were filtered. Since we sequenced a female panda, the SRY sex determination gene, which is located on the chromosome Y, was excluded in the comparison. Blat was used to align the genes.

GenBank ID	Length (bp)	% bases covered by all pieces	% bases covered by single best piece	Description
gi|145321002|gb|EF410079.1|	602	100.0	100.0	troponin C slow type (TNNC1)
gi|14669796|gb|AF395535.1|AF395535	1,938	100.0	100.0	growth hormone receptor precursor (GHR)
gi|148575268|gb|EF543744.1|	636	100.0	100.0	interleukin 6
gi|148829007|gb|EF464647.1|	681	100.0	100.0	CD9
gi|156857630|gb|EF631972.1|	448	100.0	100.0	acidic ribosomal phosphoprotein P1 (RPLP1)
gi|156857632|gb|EF631973.1|	442	98.2	53.2	ribosomal protein S15 (RPS15)
gi|163636636|gb|EU162660.1|	1,068	100.0	100.0	MHC class I antigen (Aime-1906)
gi|164507150|gb|EU195807.1|	469	100.0	100.0	ribosomal protein S19 (RPS19)
gi|167030881|gb|EU375448.1|	360	92.2	92.2	ghrelin
gi|167030885|gb|EU375450.1|	494	96.8	96.8	interleukin 15 (IL15)
gi|167030887|gb|EU375451.1|	513	100.0	100.0	leptin (ob), alternatively spliced
gi|17646741|gb|AF448453.1|	363	100.0	100.0	glycoprotein hormone common alpha subunit precursor
gi|17646743|gb|AF448454.1|	390	100.0	100.0	follicle stimulating hormone beta subunit precursor
gi|17646745|gb|AF448455.1|	426	100.0	100.0	luteinizing hormone beta subunit precursor
gi|25992712|gb|AF540936.1|	660	100.0	100.0	growth hormone precursor
gi|26516883|gb|AY161285.1|	690	100.0	100.0	prolaction precursor
gi|37551416|gb|AY327449.1|	795	96.4	96.4	prion protein

gi|4204877|gb|U56638.1|AMU 56638	744	100.0	100.0	brain derived neurotrophic factor (BDNF) gene
gi|48995460|gb|AY369779.2|	521	100.0	100.0	insulin-like growth factor I precursor
gi|54112058|gb|AY753985.1|	645	100.0	100.0	hemoglobin beta
gi|56117702|gb|AY823739.1|	468	100.0	100.0	interleukin 2 (IL2)
gi|73920576|gb|DQ166512.1|	399	100.0	100.0	interleukin 4 (IL4)
gi|85679843|gb|DQ349120.1|	335	100.0	100.0	ubiquinone-binding protein
gi|85822765|gb|DQ355514.1|	475	95.6	95.6	ribosomal protein L26
gi|88810128|gb|DQ392967.1|	564	100.0	100.0	interferon alpha 1 (IFNA1) gene
gi|89888755|gb|DQ010029.2|	501	100.0	100.0	gamma interferon (IFN-gamma)
Sum	**15,627**	**99.3**	**98.1**	

Table S5 | Comparison of assembled scaffolds and independently finished 9 BACs of the panda genome. The scaffolds were aligned with the BACs using Blast (98% identity). The alignment blocks were then chained along the BACs by a in-house program and also with manual confirmation. The overall contig coverage, single-base difference, as well as insertion/deletion difference between BAC and scaffolds were calculated. Note that the identified heterozygous SNPs and Indels were excluded previously. The estimation was performed on both the final contig and initial contigs, respectively.

BAC ID	Length (bp)	Coverage by contigs (%)	Rate of single-base difference (%) †	Median read depth on scaffold †	Median Phred score on BAC †	# of insertion and deletion ‡	# of Sanger assembly error ‡	# of hetero-zygous Indels ‡
Estimation on the final contigs								
gpbaaa	87,808	99.71	0.03	32	51	3	1	2
gpbaab	94,868	93.10	0.08	90	52	12	3	9
gpbaac	104,552	99.89	0.03	90	40	22	1	21
gpbaad	85,777	99.98	0.05	13	61	9	5	4
gpbaae	101,584	99.38	0.11	19	63	19	6	13
gpbaaf	93,450	97.08	0.12	32	47	25	8	17
gpbaag	117,932	96.85	0.12	90	38	14	6	8
gpbaah	95,133	98.49	0.04	21	55	20	11	9
gpbaak	98,323	97.83	0.07	60	24	24	12	12
Average	97,714	98.04	0.07	50	48	16	6	11
Estimation on the initial contigs								
gpbaaa	87,808	94.11	0.03	25	54	2	1	1
gpbaab	94,868	87.27	0.01	90	41	6	3	3
gpbaac	104,552	94.62	0.01	88	55	10	0	10
gpbaad	85,777	94.57	0.03	11	64	4	4	0
gpbaae	101,584	91.01	0.06	13	75	5	4	1
gpbaaf	93,450	91.15	0.07	19	54	14	7	7
gpbaag	117,932	90.75	0.01	16	61	7	6	1
gpbaah	95,133	92.39	0.02	12	55	13	10	3
gpbaak	98,323	95.69	0.03	78	36	17	10	7
Average	97,714	92.40	0.03	39	55	9	5	4

† The single-base differences were defined as mutation-like differences. To investigate the reasons, we calculated the median sequencing depth of the different bases on the scaffolds and also calculated the median Phred score of the different bases on the BACs.

‡ The insertion and deletion differences were further divided into 2 types: Sanger assembly errors and unidentified heterozygous indels. We mapped the reads onto the scaffolds and BACs in the same way, gap un-tolerated and paired-end mapping, and then assigned each insertion/deletion to the most probable type by manual checking. Note that the unidentified heterozygous indels may also contain a very small fraction

of short-read assembly errors, which were difficult to differentiate with the available data.

Table S6 | Percentage of the panda genome masked as each class of transposable elements.

	Repbase TEs		TE protiens		RepeatModeler		Combined	
	Length (Mp)	% genome	Length (Mp)	% genome	Length (Mp)	% genome	Length (Mp)	% genome
DNA	71.7	3.2	5.3	0.2	40.0	1.8	73.0	3.3
LINE	408.0	18.2	175.5	7.8	367.9	16.4	461.1	20.5
LTR	124.7	5.6	7.9	0.4	81.6	3.6	127.6	5.7
SINE	176.8	7.9	0.0	0.0	137.8	6.1	180.3	8.0
Other	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Unknown	1.4	0.1	0.0	0.0	3.1	0.1	4.4	0.2
Total	779.8	34.7	188.7	8.4	621.0	27.7	813.5	36.2

Table S7 | Comparison of abundance of each class of TEs in the panda, dog, and human genome. The same Repbase library (version 2008-08-01) was used for the panda, dog, and human genome.

	Panda		Dog		Human	
	Length (M bp)	% genome	Length (M bp)	% genome	Length (M bp)	% genome
DNA	71.7	3.2	70.0	2.9	110.4	3.9
LINE	408.0	18.2	437.6	18.3	560.3	19.7
LTR	124.7	5.6	120.2	5.0	267.0	9.4
SINE	176.8	7.9	236.4	9.9	371.7	13.1
Other	0.0	0.0	0.0	0.0	13.6	0.5
Unknown	1.4	0.1	1.3	0.1	4.6	0.2
Total	779.8	34.7	860.7	36.1	1312.2	46.1

Table S8 | Statistics of segmental duplication. "# block" for WGAC method refers to the number of pair-wise alignments, while "# block" for WSSD method refers to the number of high-depth segments.

	Panda (WGAC)			Panda (WSSD)			Dog (WGAC)		
Cutoff	# block	Median size (bp)	Genome coverage (Mb)	# block	Median size (bp)	Genome coverage (Mb)	# block	Median size (bp)	Genome coverage (Mb)
>1Kb	3095	1,657	10.4	5,485	1,696	34.3	11,516	1,756	43.8
>5Kb	164	6,897	1.9	525	6,556	11.5	1,641	7,991	30.3
>10Kb	38	13,363	0.7	89	13,033	3.8	618	17,411	23.2
>50Kb	1	114,485	0.1	0	0	0.0	88	99,170	12.0

Table S9 | Syntenic regions between panda and dog, and between panda and human. BlastZ was used to align the genomes, and the "net" output results that allow gaps and local small rearrangements were used to define large-scale syntenic regions.

	Query aligned (bp)	% of query	Target aligned (bp)	% of target	# blocks
Panda vs dog	2,224,446,488	96.74	2,271,445,649	92.90	13,263
Panda vs human	2,191,353,032	95.3	2,662,329,982	88.08	10,896

Table S10 | Intra-chromosomal and inter-chromosomal rearrangement between panda and dog. Syntenic segment is defined as continuous regions without any order or orientation change. Different cutoffs of syntenic segment, from 5Kb to 500Kb, were used to find inter-chromosomal and intra-chromosomal breakpoints. The human genome was used as an outgroup to differentiate whether panda or dog changed.

	Intra-chromosomal rearrangement				Inter-chromosomal rearrangement			
Cutoff (Kb)	# scaffold	# breakpoint	# panda changed	# dog changed	# scaffold	# breakpoint	# panda changed	# dog changed
5	268	468	89	291	59	59	13	46
10	154	254	42	186	59	59	13	46
20	79	123	15	101	56	56	12	44
50	31	41	7	34	53	53	11	42
100	18	20	4	16	42	42	7	35
200	7	7	2	5	33	33	6	27
500	2	2	1	1	17	17	3	14

Table S11 | Statistics of homology-based gene predictions. "With synteny" means genes predicted on regions with synteny to the human or dog, and the fragmental genes were conjoined by building gene-scaffolds. "Out of synteny" means genes predicted on regions without synteny evidence to the other species; Pseudo-genes, are those containing more frame errors than a specified threshold.

	With synteny (complete)	With synteny (fragmental)	Out of synteny	Pseudo-genes	Final prediction
Human projections	16,412	166	2,725	2,958	19,303
Dog projections	16,557	145	2,543	2,422	19,245

Table S12 | General statistics of each gene set and integrated prediction. Gene length included the exon and intron regions but excluded UTRs.

Gene sets	Total genes	Average gene length (bp)	Average CDS length (bp)	Average CDS GC Ratio	Average Exons per gene	Average Exon length (bp)	Average Intron length (bp)
Human projections	19,303	26,571	1,497	0.53	8.5	176	3,344
Dog projections	19,245	24,369	1,469	0.53	8.5	172	3,042
Genscan predictions	44,428	33,216	1,253	0.53	7.8	161	4,732
Augustus predictions	29,238	19,072	1,097	0.57	7.1	177	3,472
Integrated prediction	21,001	26,857	1,479	0.53	8.4	175	3,510

Table S13 | Enrichment of PAHTER terms in segmental duplication (SD). The gene enrichment in SD regions was compared to all the reference genes (Ref), and P-values were calculated by Fisher-exact test.

PANTHER class	PANTHER iterm	# Ref genes	# SD genes	P-value
Biological Process	B-cell- and antibody-mediated immunity	214	61	2.66E-34
Biological Process	Chemosensory perception	435	76	8.27E-30
Biological Process	Biological process unclassified	4,035	198	1.04E-08
Biological Process	Cell surface receptor mediated signal transduction	1,797	99	2.27E-07
Biological Process	T-cell mediated immunity	193	14	5.33E-03
Biological Process	Fertilization	37	5	7.20E-03
Molecular Function	Immunoglobulin	135	61	1.28E-43
Molecular Function	G-protein coupled receptor	786	81	1.31E-18

Molecular Function	Molecular function unclassified	3,725	186	7.60E-09
Molecular Function	Intermediate filament	94	8	1.00E-02
Molecular Function	Storage protein	27	4	1.20E-02
Molecular Function	Oxygenase	89	7	2.21E-02

Table S14 | Gain and loss of genes in panda.

a. Gene gain in panda

GOID	GO description	Number of genes	P-value
GO:0004872	receptor activity	29	3.93E-04
GO:0004984	olfactory receptor activity	23	8.75E-10
GO:0008083	growth factor activity	5	3.07E-02
GO:0003735	structural constituent of ribosome	5	4.37E-02
GO:0005496	steroid binding	2	4.60E-02
GO:0008466	glycogenin glucosyltransferase activity	2	8.24E-04
GO:0005542	folic acid binding	2	8.52E-03
GO:0003954	NADH dehydrogenase activity	2	1.32E-02

b. Gene loss in panda

GOID	GO description	Number of genes	P-value
GO:0004872	receptor activity	78	1.37E-09
GO:0004984	olfactory receptor activity	61	9.67E-25
GO:0001664	G-protein-coupled receptor binding	3	3.95E-03
GO:0003729	mRNA binding	3	4.01E-02
GO:0004499	flavin-containing monooxygenase activity	2	1.78E-02
GO:0008191	metalloendopeptidase inhibitor activity	2	3.60E-02
GO:0004029	aldehyde dehydrogenase (NAD) activity	2	2.94E-02
GO:0005006	epidermal growth factor receptor activity	2	2.77E-03

Table S15 | Selected PANTHER categories over-represented among genes predicted to be under positive selection. Shown are numbers of PSGs and of all genes available assigned to each category or a descendant category, and one-sided (nominal) *P-values* from the Mann-Whitney U (MWU) and Fisher's exact (FE) tests. Note that the MWU *P-values* do not consider whether or not each gene is predicted to be a PSG, but instead indicate the degree to which the LRT P -values for the genes of each category are shifted toward small values. Consequently, classes of genes experiencing relaxation of constraint but not positive selection may obtain small MWU *P-values*. In contrast, the FE *P-values* indicate over-representation of the identified PSGs within each category (or, equivalently, over-representation of each category among the PSGs). Bold indicates significance after a conservative correction for multiple testing (FWER, 0.05, Holm correction[40]).

Category	Description	Gene number		*P-value*	*P-value*
		PSGs	All	MWU	FE
		Panda			
BP00155	Macrophage-mediated immunity	5	70	1.63 x 10^{-7}	0.0013
BP00210	Blood circulation and gas exchange activity	1	11	4.44 x 10^{-5}	0.1221
BP00148	Immunity and defense	20	735	5.19x 10^{-5}	0.0002
BP00255	Cytokine/chemokine mediated immunity	4	73	5.73 x 10^{-5}	0.0095
MF00173	Defense/immunity protein	11	153	4.35 x 10^{-5}	1.54 x 10^{-6}
MF00018	Chemokine	3	25	4.19 x 10^{-5}	0.0033
		Dog			
BP00120	Cell adhesion-mediated signalling	14	211	5.33 x 10^{-2}	1.32 x 10^{-9}
BP00274	Cell communication	16	703	6.97 x 10^{-2}	8.86 x 10^{-5}
BP00148	Immunity and defense	11	744	7.78 x 10^{-6}	0.0269
BP00155	Macrophage-mediated immunity	1	74	9.18 x 10^{-6}	0.4352
MF00259	Cadherin	7	45	6.00 x 10^{-2}	6.34 x 10^{-7}
MF00040	Cell adhesion molecule	11	204	1.00 x 10^{-2}	1.15 x 10^{-7}
MF00173	Defense/immunity protein	4	160	1.11 x 10^{-5}	0.0357
MF00017	Cytokine	1	64	3.94 x 10^{-5}	0.3904

Table S16 | Selected GO categories over-represented among genes predicted to be under positive selection.

Category	Description	Gene numbers PSGs	All	*P-value* MWU	*P-value* FE
		Panda			
GO:0002526	Acute inflammatory response	9	44	3.55×10^{-3}	4.48×10^{-9}
GO:0005576	Extracellular region	33	1160	8.76×10^{-9}	2.88×10^{-7}
GO:0006953	Acute-phase response	5	14	3.58×10^{-3}	1.40×10^{-6}
GO:0006955	Immune response	16	355	4.04×10^{-5}	2.38×10^{-6}
GO:0045087	Innate immune response	8	77	1.09×10^{-3}	3.56×10^{-6}
GO:0009611	Response to wounding	13	292	1.48×10^{-8}	2.52×10^{-5}
GO:0006952	Defense response	16	341	2.44×10^{-7}	1.44×10^{-6}
GO:0007596	Blood coagulation	4	61	5.28×10^{-6}	0.0053
GO:0007599	Hemostasis	4	66	6.85×10^{-6}	0.0069
GO:0004872	Receptor activity	16	821	1.49×10^{-5}	0.0181
		Dog			
GO:0019835	Cytolysis	5	8	2.84×10^{-2}	2.73×10^{-8}
GO:0016337	Cell-cell adhesion	10	149	3.94×10^{-2}	3.19×10^{-7}

Table S17 | Numbers of reads aligned onto the assembled scaffold sequences. U0, U1, U2, means reads with single best mapping locations and 0, 1, or 2 mismatches, respectively. While R0, R1, and R2 means reads with multiple equal best mapping locations.

	# reads	% reads	# bases	% bases
Unique	2,580,473,364	76.36	131,985,139,571	74.98
U0	1,974,598,597	58.43	98,360,142,747	55.88
U1	426,385,995	12.62	23,603,028,493	13.41
U2	179,488,772	5.31	10,021,968,331	5.69
Repeat	224,928,958	6.66	10,878,756,779	6.18
R0	66,347,893	1.96	3,099,072,464	1.76
R1	62,967,350	1.86	3,106,793,043	1.77
R2	95,613,715	2.83	4,672,891,272	2.65
Others	74,440,445	2.20	3,421,402,849	1.94
Total	2,879,842,767	85.22	146,285,299,199	83.11

Table S18 | Substitution matrix of the panda heterozygous SNPs in the whole genome. The ratio of transition / transversion is 2.1.

	A	C	G	T
A	-	-	-	-
C	230,535	-	-	-
G	909,027	190,142	-	-
T	216,029	908,331	228,285	-

Table S19 | Statistics of heterozygous Indels in the whole genome.

Indel size	**Genome**		**CDS**	
	Number	**Rate (*1E-4)**	**Number**	**Rate (*1E-4)**
1-bp	173,825	0.774	123	0.040
2-bp	46,053	0.205	56	0.018
3-bp	19,222	0.086	73	0.024
4-bp	19,479	0.087	34	0.011
5-bp	5,844	0.026	13	0.004
6-bp	3,535	0.016	18	0.006
Total	267,958	1.193	317	0.103

Table S20 | Statistics of heterozygous structural variations in the whole genome.

	SV #	Median length (bp)	% overlap TEs
Indel(>100bp)	4359	150	71.5
Inversion	20	356	90.0

Supplementary results

Additional panda, dog, and human repeat comparison

The content of most TE classes was similar between panda and dog. Panda contained slightly more LTR retrotransposons than dog (5.6% vs 5.0%), but fewer SINEs (Short Interspersed Elements) (7.9% vs 9.9%). (Supplementary Table 6, 7) There are about 500 Mb fewer TE sequences in the panda genome than in the human genome, which is likely to be the main reason for the genome size difference (2.4 Gb vs 3.0 Gb).

We analyzed the divergence rate of the TE elements in the panda genome by using both the Repbase and the *RepeatModeler* TE libraries. Nearly all the identified panda TE copies had a >10% divergence rate from the consensus in Repbase. This high divergence rate may be related to the fact that the Repbase TE consensus sequences were annotated using mammalian genomes other than the panda. Using *RepeatModeler* TEs, we found that about 70 Mb of TE sequences (3% of the genome) had a <10% divergence rate from the consensus (Supplementary Fig. 9), which are likely to be active TEs of recent origin. These include SINE/Lys, which are a family of carnivore-specific SINEs that are thought to be derived from transfer RNA (tRNA)-Lys[41], and young L1 elements that are LINEs derived from the youngest mammalian-wide L1MA element in Carnivora lineage[41]. Studies have shown that these two TE families are active in dog[41] and cat[42].

We found that SINE/Lys TEs comprised 5.5% of the panda genome, and of these, the youngest SINE/Lys subfamily had 150,280 copies and covered 1.3% of the genome. Despite its high abundance, the average divergence of these SINE/Lys elements from the consensus sequence was only 5.6%. This provides evidence that these are recent insertions in the panda genome that occurred after speciation from the dog. Within the SINE/Lys family, SINEC_b1 and SINEC_b2 were 2.5 times more abundant in the panda than in the dog genome, while SINEC_a1, SINEC_a2, SINE_Cf, SINC_Cf2 and SINEC_Cf3 were 14.3 times more abundant in the dog than in the panda genome.

Similar to the SINE/Lys TEs, carnivore-specific L1 elements comprised 5.1% of the panda genome. The youngest L1 subfamily had 18,467 copies in the panda, covered 0.9% of the genome, and had an average divergence rate of about 4.7% from their consensus sequence. Among the LINE/L1 family, L1_Canid subfamily was 2.2 times more abundant in the panda genome; while L1_Canis subfamily and L1_Cf were 106.6 times more abundant in the dog genome (Figure SA1). Although the biological effect of these new SINE and LINE insertions is unknown, their high activity may have an impact on panda diseases and its genome diversity.

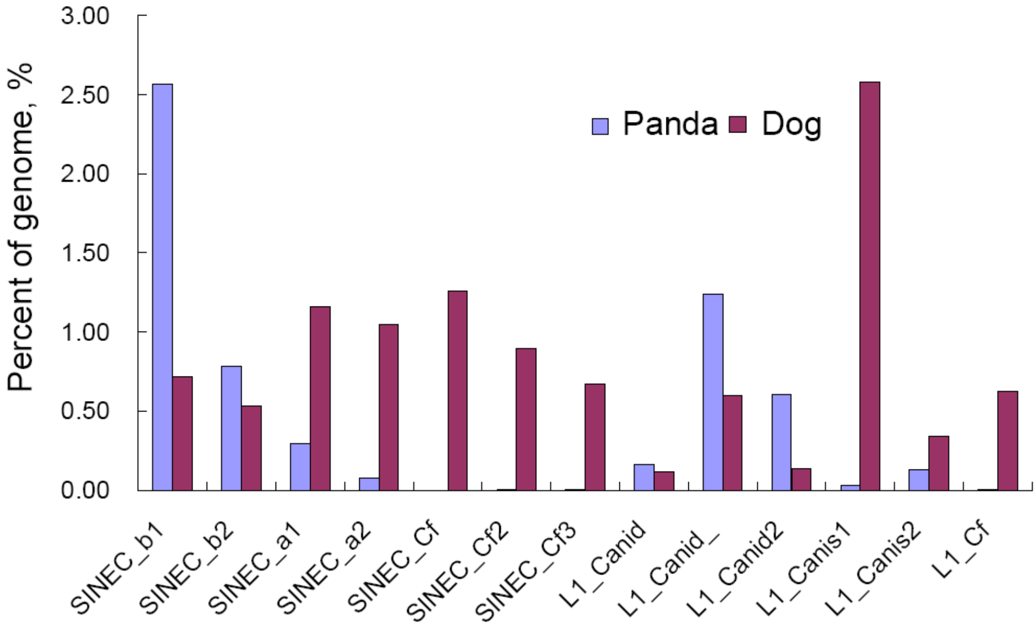


Figure SA1 | Enriched carnivore-specific SINE/Lys and LINE/L1 TE families in the panda and dog genome. The consensus sequences of the two TE families in the Repbase database were used to identify the elements in the panda and dog genomes using RepeatMasker.

Pseudogenes

We identified a total of 1,537 processed pseudogenes derived from retrotransposition with no introns[43]. We then compared the pseudogenes in the panda with those in the human and mouse, and found that the genes that have many retrotransposed pseudogenes in the human and mouse are also likely to have multiple pseudogenes in panda. For example, ribosomal proteins are among the genes that have the most pseudogenes, with 536 copies in the panda genome.

Non-coding RNA genes

Identifying tRNA genes in panda and other carnivore genomes was affected by the abundant number of (tRNA)-Lys derived SINE TEs. We predicted 20,807 tRNA genes and 84,440 tRNA pseudogenes by tRNAScan-SE[44], but 20,551 of these "genes" and 84,317 of "pseudogenes" were masked as SINE TEs by RepeatMasker, indicating that most of them are false positive predictions due to the abundance of tRNA-derived SINE TEs in the panda genome. After removing the SINE-masked tRNAs, we obtained a clean set of 256 tRNA genes.

To improve discrimination of functional tRNA genes, we exploited comparative genomic analysis between panda and human. In contrast to panda, the human genome contained quite few tRNA-derived SINE TEs, 519 tRNA genes were predicted by tRNAScan-SE and only 8 were masked as SINE TEs. We aligned the remaining 511 human tRNA genes to different categories of panda tRNAs, and the identity curve showed that the clean tRNA genes are much more conserved than the pseudogenes or SINE-masked tRNA genes (Figure SA2). We also observed that there are 135 SINE-masked tRNA genes having more than 95% identity to human, which are likely to be functional genes but mis-identified as SINEs. So we combined them with the clean tRNA genes to form a final set of 391 tRNA genes, of which 43 have introns. The set represents all 46 expected anticodons and none violates the wooble rules. There are 247 (63%) tRNA genes, more than a half, localized in 17 clusters defined as containing no less than 5 tRNA genes in a scaffold.

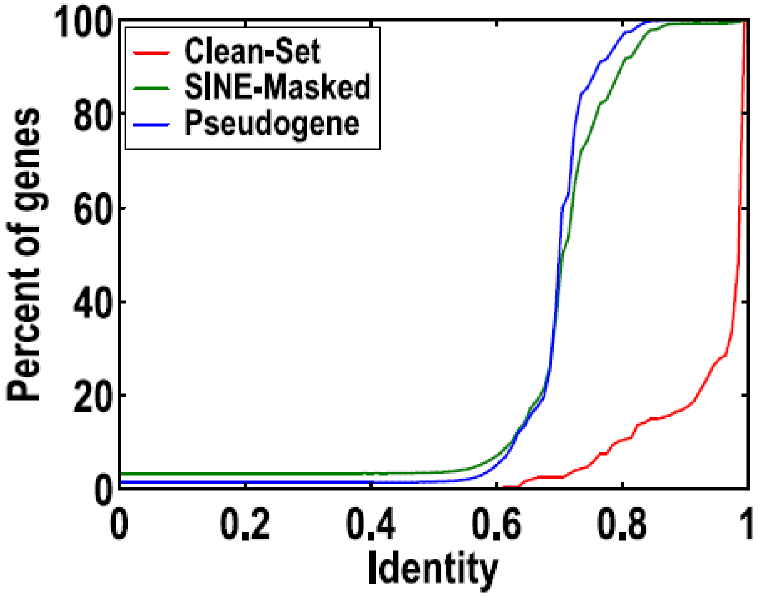


Figure SA2 | Homology of panda tRNA genes to human. The Clean-Set had a relatively much higher identity rate than the other two sets, which are mostly composed of SINE/Lys TEs.

We identified 249 rRNA fragments, by aligning the human template sequences to the panda genome. They are rather fragmental, presumably owing to assembly issues. We also analyzed the panda genome for other known classes of non-coding RNA genes. We predicted 39,970 miRNA, 328 C/D-box snoRNA, 232 H/ACA-box snoRNA, 22 scaRNA, and 939 spliceosomal RNA candidate genes, using the Rfam's method[16] with the recommend score cutoff. The prediction of miRNAs was greatly interfered by the abundant LINE/L1 TEs. Among the 39,970 miRNA predictions, 38,666 (96.7%) were masked by TEs, most (93.4%) of which were LINE/L1, especially carnivore-specific LINE/L1 families (48.7%). However, there is just 4.6% for other ncRNA types. After homology filtering against known mammalian ncRNA sequences, we identified 307 miRNA, 143 C/D-box snoRNA, 108 H/ACA-box snoRNA, 15 scaRNA, and 235 spliceosomal RNA genes.

Xist RNA is one of the most interesting ncRNAs since it is involved in X-chromosome inactivation. The location of the Xist (X inactive-specific transcript) non-coding RNA gene was determined by alignment with other species. The splice junctions were determined by examining the alignment for homologous splice sites. The total predicted length of the gene is ~24,500bp, organized into 5 exons. The exon structure is predicted to be the same as the canine gene, although the 3' end of the gene is not well conserved and so the location of the gene end is not clear (Figure SA3).

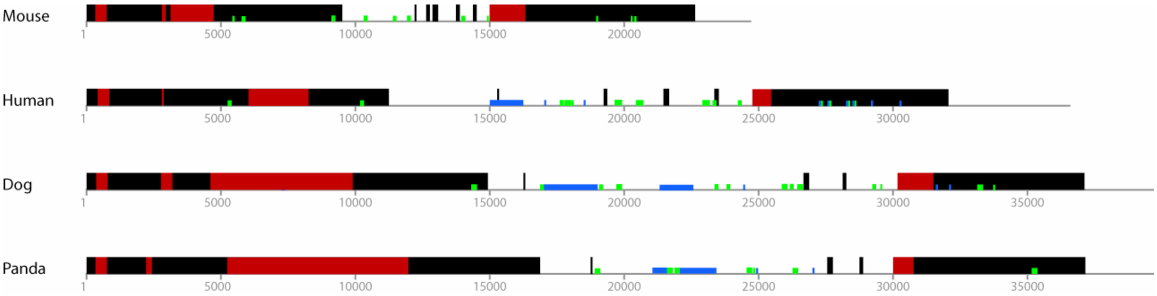


Figure SA3 | Exon structure of Xist across 4 mammalian species. Black boxes denote exons and conserved tandem repeats (red), as well as LINE (blue) and SINE (green) repetitive elements are shown.

Detection of mammalian conserved non-coding elements

We detected evolutionarily conserved elements among mammals, i.e. the discrete regions under purifying selection, using PhastCons on the 5-way genome alignments that included human, panda, dog, mouse, and rat. A total of 107.0Mb (3.8%) was identified as conserved elements on human genome, a little less than previous reports of about 5%[45,46], which may be due to the difference of our cutoffs. The counterpart on the panda genome is 93.6Mb (4.2%), and overlapped with ~80% of coding exons. The conserved non-coding elements (CNE) were obtained by excluding the CDS regions from the conserved elements, and requiring length >= 50bp. There are in total of 424,787 CNEs, with median length of 118 bp. The extent and function of the large fraction of non-coding conserved sequences remain unclear, but these sequences are likely to include regulatory elements, structural elements and RNA genes[47]. Indeed, about 50% of predicted miRNAs and snRNAs, 60% of the predicted CpGProd and EP3 promoters, and 70% of the homology annotated UTRs overlapped with the CNEs.

Regulatory elements and nucleosome positioning

We identified 50,773 CpG islands in the non-repetitive portions of panda genome, using the thresholds of region length >= 500bp, GC content >= 55%, and observed/expected ratio >= 0.65[48]. The number of CpG islands is similar to that in the dog (54,310). However, it is lower in the human (28,750) and mouse (17,238) genomes, which may be related to the higher fraction of genomes with high (G+C) content. We predicted 16,589 promoter regions using CpGProD[49], which was more than in human (15,155) and mouse (13,893), but less than dog (23,761) (Table SA1). A similar result was obtained using an alternative software EP3[50], and half of predictions overlapped between CpGProD and EP3.

Table SA1 | Number of predicted regulatory elements.

	Panda	Dog	Human	Mouse
CpG islands	50,773	54,310	28,750	17,238
CpGProD	16,589	23,761	15,155	13,893
EP3	18,717	27,555	16,976	12,101

We mapped the nucleosomes in the panda genome with a curvature profile[51]. The length of core DNA is conserved among all mammalians, whereas the length of linker DNA is variable. Spectrum of the curvature profile indicated the core DNA is 147 bp in length in the panda, the same as that for dog and cat, but that the linker DNA is 23 bp, which is shorter than that of the dog (35bp) and cat (41bp). The total length of nucleosomal DNA is 170 bp, 182 bp, and 188 bp for panda, dog, and cat, respectively. Nucleosome-free regions (NFR) near the start of ORFs show an ~80 bp shift to downstream, as compared to that in dogs (-300 bp in panda; -380 bp in dog), indicating the panda has shorter 5’ untranslated regions (5’UTR) (Figure SA4).

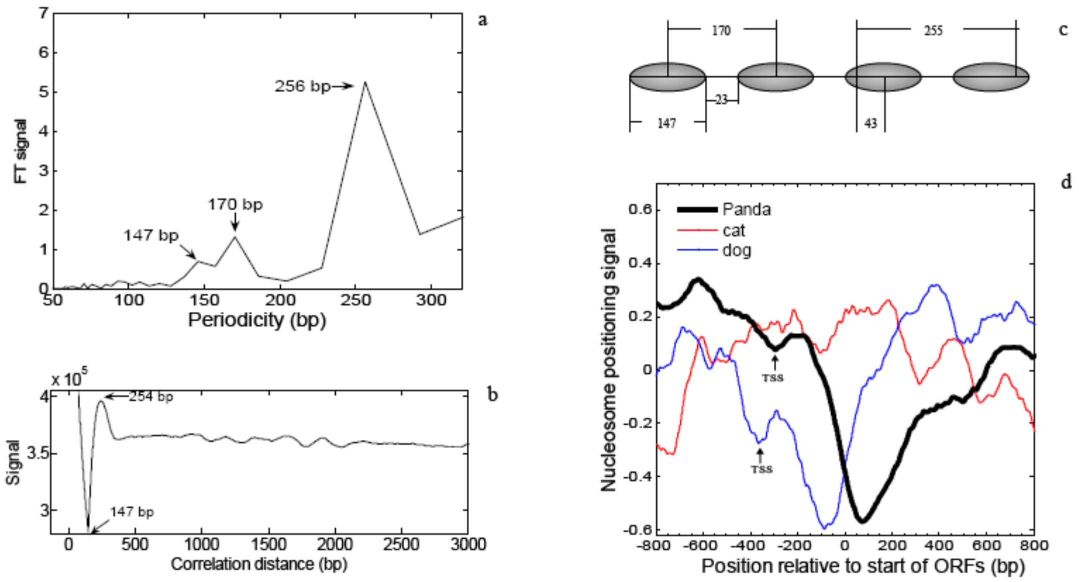


Figure SA4 | Results from nucleosome analysis in the panda genome. a, Fourier spectrum of nucleosome positioning signal (curvature profile); arrows indicate the peaks positions. The inferred length of the core DNA was 147 bp, and the distance between adjacent nucleosome midpoints was ~170 bp in panda. **b**, the autocorrelation signal of curvature profile, the first valley indicates the similarity is lowest when the nucleosome psotioning signal shifts a length of a nucleosome (147 bp) to itself. The 256-bp peak in spectrum and the 254-bp peak in autocorrelation signal reflect the distance from the start of a nucleosome to its immediate neighbor's end (see c). **c**, the inferred spacing structure of nucleosomes in panda. **d**. nucleosome distribution in the vicinity of the start of 1,274 open reading frames (ORFs), the position of the first trough upstream of the start codon is identified as transcription start site (TSS) (indicated by arrows).

Neutral substitution rate estimation using ancestral repeats

Ancestral repeat is a powerful source for unbiased neutral substitution rate estimation since its high copy number and broad genomic distribution[41,45]. Evolving from the common ancestor of human, cat, dog and panda, the nucleotide divergence in these four genomes is similar with a panda to dog ratio at 0.97, a panda to cat ratio at 0.99 and a panda to human ratio at 1.03. The cause of the small differences of nucleotide divergence in these species is unknown, but it may be partly explained by their different generation time and metabolic rates[45]. When we only consider nucleotide change after the separation from the common ancestor of dog and panda, the nucleotide substitution rate is even lower in panda than that in dog (panda to dog ratio at 0.90).

Phylogenetic analyses

We constructed a phylogenetic tree of the panda and the other sequenced mammalian genomes using the 7,034 single-copy orthologuous genes and 4-fold degenerate sites (Figure SA5). We found that the neutral divergence rate of the panda (Ursidae) lineage is slightly smaller than that of the dog (Canidae) lineage after speciation from their common ancestor (0.09 vs 0.10, substitution per site). The slower molecular clock might be explained by the body size hypothesis[52] or the generation-time hypothesis[53], which propose that the larger the body size is or the longer the generation-time is, the slower the molecular clock. Interestingly, the dN/dS of the panda lineage is about the same as that of the dog lineage (~0.18 for both) and smaller than that of the human lineage (0.24). According to Ohta's nearly neutral theory[54], weak selection can lead to different evolutionary fates due to different population sizes and reduced population size can increase the fixation of slightly deleterious mutations. This indicates that the average population size history in the Ursidae lineage and Canidae lineage may have been similar since divergence from their last common ancestor.

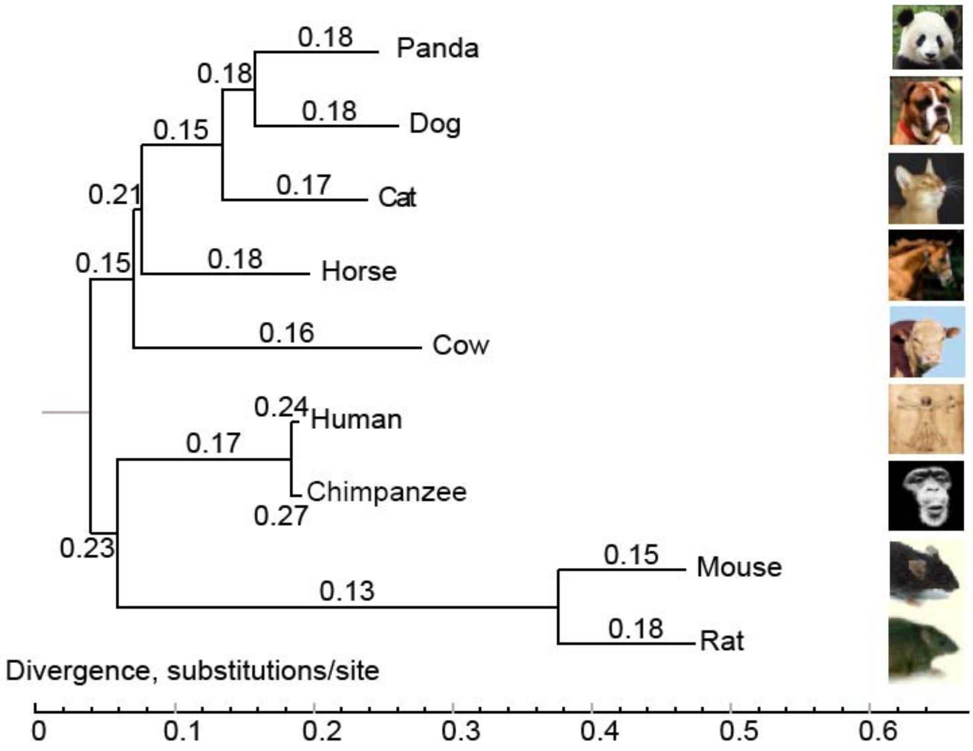


Figure SA5 | Phylogenetic tree constructed with 7,034 single-copy gene families on 4-fold degenerate sites. The branch length represents the neutral divergence rate. Numbers shown on the branch represent the dN/dS ratio on that branch. The posterior probabilities (credibility of the topology) for inner nodes are all 100%.

We estimated time of divergence for the panda by incorporating the 20 ortholog genes from the American black bear, dog, human, mouse, and opossum (as an outgroup). The estimated divergence time between panda and bear is 14 Mya, which is concordant with previous estimates[55] (Figure SA6). Assuming that the generation time for panda is 4.5–6.5 years[56], the substitution rate in panda would be 1.3e-9 substitutions per site per year or about 0.6e-8~0.8e-8 substitutions per site per generation.

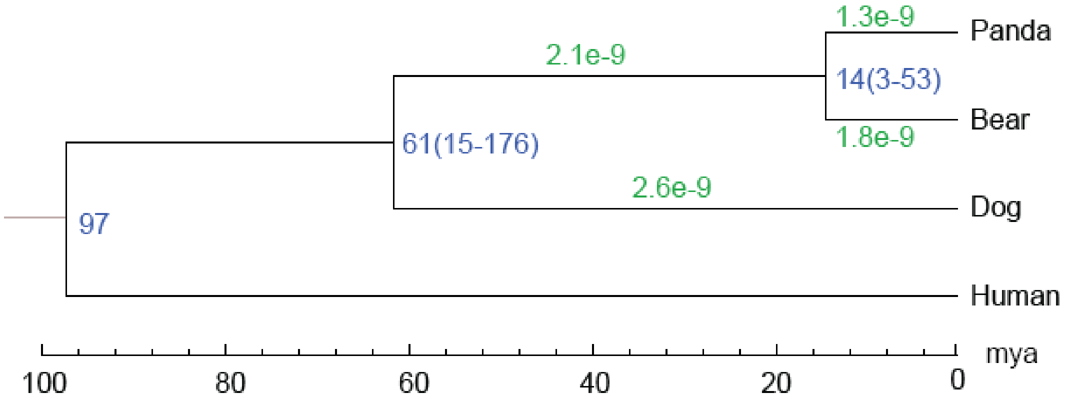


Figure SA6 | Estimation of divergence time and substitution rate. The green numbers on the branches are the estimated substitution rate (substitutions per site per year). The blue numbers on the nodes are the divergence time from present (million years ago, Mya). The calibration time (97 Mya) from human-dog divergence was derived from the TimeTree database (http://www.timetree.org).

Analysis of olfactory receptor genes

We examined several of the nasal chemoreceptor gene families (Olfactory receptors, OR; Trace amine-associated receptor, TAAR; Vomernasal receptor I, V1R; Vomernasal receptor II, V2R), which may also play a role in developing dietary behaviours. The panda OR repertoire had 659 putative intact genes, which is less than the 811 putative intact ORs in the dog[57]. We did not identify any V2R members in the panda genome when using a requirement for intact open reading frames (ORFs); likewise, no V2R members have been identified in the dog genome[58]. We did identify 13 putative intact V1Rs and 11 putative intact TAARs in the panda, which is more than that in the dog (8 intact V1Rs[59] and 2 intact TAARs[60]).

Analysis of pigmentation related genes

To gain some insight into the distinctive black and white pattern of panda skin, we investigated the pigmentation genes. We identified 50 pigmentation genes in the panda genome and compared them to those in other mammals[61] and found that 5 of these genes show positive selection in the panda lineage: the breast cancer 1 gene (*Brca1*), keratin 2 (*Krt2*), SRY-box containing gene 10 (*Sox10*), tyrosinase (*Tyr*), and melanophilin (*Mlph*) (Table SA2). *Brca1*, *Krt2* and *Sox10* are genes known to be involved in melanocyte development and differentiation in the house mouse[62]. *Tyr* is a melanosomal enzyme, which converts tyrosine to melanin[63]. *Mlph* is a Rab-effector protein involved in melanosome transport to the actin cytoskeleton in melanocytes[64]. In humans, mutations in these genes are associated with diseases that are characterized by de-pigmentation or abnormal pigment distribution of the hair and skin, such as Waardenburg syndrome[65]. Future experimental analyses involving *in vitro* or gene expression studies of these five positively selected pigmentation genes may shed light on the unique pigmentation pattern of the giant panda.

Table SA2 | Pigmentation genes under positive selection in panda lineage.

Genes	Ln L (ω =1)	Ln L (ω >1)	2Δ	P
Tyr	-4431.01	-4427.24	7.53	0.0061
Brca1	-11789.86	-11786.58	6.56	0.0104
Krt2	-4974.79	-4968.63	12.32	0.0004
Mlph	-4448.87	-4444.58	8.59	0.0034
Sox10	-2911.19	-2908.32	5.75	0.0165

"Pseudo-thumb" and Homeobox gene family

Three gene families related to the development of limb were identified from the panda genome, including posterior Hox gene family[66,67], Fibroblast growth factor (FGF) family[68,69], and bone morphogenetic protein (BMP) gene family[70]. The development of limb keeps conserved in vertebrates, which is a delicately controlled process in both temporal and spatial scale[71]. Consistent with this, evolutionary analysis indicated that the three gene families changed slightly between the Order Carnivora and the Order Primata, including panda, dog and human. The posterior Hox gene family, including Hox9-13[67], maintained conservation among panda, dog, and human genomes in both gene order and coding DNA sequence, and even in introns for some members. But these genes vary substantially in putative transcription regulation region and some introns. This indicated that these genes have different transcriptional patterns, which might be related to the morphological diversity of the forelimb among panda, dog and human.

Both BMP and FGF gene families play essential roles in limb development, particularly the in the formation of the limb bud[70]. In the panda genome we identified 22 members of FGF family and 10 members of BMP family, which is the number as that in the dog and human. Those genes reported to be related to limb bud formation, such as FGF4[72], FGF8[72], FGF10[68], BMP2[70] and BMP7[73], maintained conservation in the coding sequences between panda and dog. To analyze the transcription region of these genes, we compared the 2 kb upstream genome sequence of the first exon between panda and dog. Because only the coding DNA sequence was marked in the gene annotation data, we used the start codon as the start of the first exon to simplify the analysis. Generally these regions include transcription factor combined region and 5'-untranslated region, both of which contribute to transcription regulation. The result showed that these genes vary substantially in these regions. This may indicate that the transcription regulation of these genes differs between dog and panda, which may contribute to, at least in part, to the panda's distinct feature of a pseudo-thumb.

References

1. Li, R., ZHu, H. & Wang, J. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res (in press)* (2009).
2. Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* **98**, 9748-53 (2001).
3. Smit AFA, H.R., Green, P. RepeatMasker Open-3.0. *1996-2004* (http://www.RepeatMasker.org.).
4. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-7 (2005).
5. Bao, Z. & Eddy, S.R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**, 1269-76 (2002).
6. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-8 (2005).
7. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
8. She, R., Chu, J.S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res* **19**, 143-9 (2009).
9. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-95 (2004).
10. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10**, 516-22 (2000).
11. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215-25 (2003).
12. Mulder, N.J. et al. New developments in the InterPro database. *Nucleic Acids Res* **35**, D224-8 (2007).
13. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
14. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
15. Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335-7 (2009).
16. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121-4 (2005).
17. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**, D154-8 (2008).
18. Lestrade, L. & Weber, M.J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* **34**, D158-62 (2006).
19. Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103-7 (2003).
20. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-15 (2004).
21. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).
22. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**, 1005-17 (2001).

23. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-4 (2008).
24. Li, H. et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34**, D572-80 (2006).
25. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7 (2004).
26. Posada, D. & Crandall, K.A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-8 (1998).
27. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755 (2001).
28. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
29. Hedges, S.B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971-2972 (2006).
30. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**, 568-73 (1998).
31. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-91 (2007).
32. Kosiol C, V.T., da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. Patterns of positive selection in six mammalian genomes. *PLoS Genetics* **4(8): e1000144. doi:10.1371/journal.pgen.1000144.**(2008).
33. Labarga A, V.F., Anderson M, Lopez R. Web services at the European Bioinformatics Institute. *Nucleic Acids Research* **35**, W6-11 (2007).
34. Posada D, C.K. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818 (1998).
35. DL, S. *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*, (Sunderland, MA, 1998).
36. Arnason U, A.J., Gullberg A, Harley EH, Janke A, Kullberg M. Mitogenomic relationships of placental mammals and molecular estimates of their divergences. *Gene* **421**, 37–51 (2008).
37. Z, Y. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
38. Yang Z, W.W., Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**, 1107–1118 (2005).
39. Li, R. et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res* **19**, 1124-32 (2009).
40. S, H. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979).
41. Kirkness, E.F. et al. The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898-903 (2003).
42. Pontius, J.U. et al. Initial sequence and comparative analysis of the cat genome. *Genome Res* **17**, 1675-89 (2007).
43. Zhang, Z., Harrison, P.M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13**, 2541-58 (2003).

44. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-64 (1997).

45. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-19 (2005).

46. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).

47. Kim, S.Y. & Pritchard, J.K. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet* **3**, 1572-86 (2007).

48. Takai, D. & Jones, P.A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* **99**, 3740-5 (2002).

49. Ponger, L. & Mouchiroud, D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* **18**, 631-3 (2002).

50. Thomas Abeel, Y.S., Eric Bonnet, et al. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.* **18**, 310-323 (2008).

51. Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **442**, 772-8 (2006).

52. Martin, A.P. & Palumbi, S.R. Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A* **90**, 4087-91 (1993).

53. Li, W.H., Ellsworth, D.L., Krushkal, J., Chang, B.H. & Hewett-Emmett, D. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol* **5**, 182-7 (1996).

54. Ohta, T. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics* **23**, 263 (1992).

55. Krause, J. et al. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evol Biol* **8**, 220 (2008).

56. Pan W, L.Z., Zhu X, Wang D, Wang H, Long Y, Fu L, Zhu X. A chance for lasting survival. *Beijing: Beijing University Press.* (2001).

57. Niimura, Y. & Nei, M. Extensive gains and losses of olfactory receptor genes in Mammalian evolution. *PLoS One* **2**, e708 (2007).

58. Shi, P. & Zhang, J. Comparative genomic analysis identifies an evolutionary shift of vomeronasal receptor gene repertoires in the vertebrate transition from water to land. *Genome Res* **17**, 166-74 (2007).

59. Grus, W.E., Shi, P., Zhang, Y.P. & Zhang, J. Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals. *Proc Natl Acad Sci U S A* **102**, 5767-72 (2005).

60. Grus, W.E., Shi, P. & Zhang, J. Largest vertebrate vomeronasal type 1 receptor gene repertoire in the semiaquatic platypus. *Mol Biol Evol* **24**, 2153-7 (2007).

61. Kosiol, C. et al. Patterns of positive selection in six Mammalian genomes. *PLoS Genet* **4**, e1000144 (2008).

62. Bennett, D.C. & Lamoreux, M.L. The color loci of mice--a genetic century. *Pigment Cell Res* **16**, 333-44 (2003).

63. Kwon, B.S., Wakulchik, M., Haq, A.K., Halaban, R. & Kestler, D. Sequence analysis of mouse tyrosinase cDNA and the effect of melanotropin on its gene expression. *Biochem Biophys Res Commun* **153**, 1301-9 (1988).

64. Matesic, L.E. et al. Mutations in Mlph, encoding a member of the Rab effector family, cause the melanosome transport defects observed in leaden mice. *Proc Natl Acad Sci U S A* **98**, 10238-43 (2001).

65. Waardenburg, P.J. A new syndrome combining developmental anomalies of the eyelids, eyebrows and nose root with pigmentary defects of the iris and head hair and with congenital deafness. *Am J Hum Genet* **3**, 195-253 (1951).

66. Capdevila, J. & Izpisua Belmonte, J.C. Patterning mechanisms controlling vertebrate limb development. *Annu Rev Cell Dev Biol* **17**, 87-132 (2001).

67. Maconochie, M., Nonchev, S., Morrison, A. & Krumlauf, R. Paralogous Hox genes: function and regulation. *Annu Rev Genet* **30**, 529-56 (1996).

68. Sekine, K., Ohuchi, H., Fujiwara, M., Yamasaki, M., Yoshizawa, T., Sato, T., Yagishita, N., Matsui, D., Koga, Y., Itoh, N. & Kato, S. Fgf10 is essential for limb and lung formation. *Nat. Genet.* **21**, 138-141 (1999).

69. Lewandoski, M., Sun, X. & Martin, G.R. Fgf8 signalling from the AER is essential for normal limb development. *Nat Genet.* **26**, 460-463 (2000).

70. Niswander L, M.G. FGF-4 and BMP-2 have opposite effects on limb growth. *Nature* **361**, 68-71 (1993).

71. Johnson, R.L. & Tabin, C.J. Molecular models for vertebrate limb development. *Cell* **90**, 979-90 (1997).

72. Boulet, A.M., Moon, A.M., Arenkiel, B.R. & Capecchi, M.R. The roles of Fgf4 and Fgf8 in limb bud initiation and outgrowth. *Dev Biol* **273**, 361-72 (2004).

73. Merino, R. et al. The BMP antagonist Gremlin regulates outgrowth, chondrogenesis and programmed cell death in the developing limb. *Development* **126**, 5515-22 (1999).