

Genomics & Bioinformatics

Vol 463 | 21 January 2010 | doi:10.1038/nature08696

nature

Journal Club

ARTICLES

The sequence and *de novo* assembly of the giant panda genome

BIOL 497, 597 – Special Topics

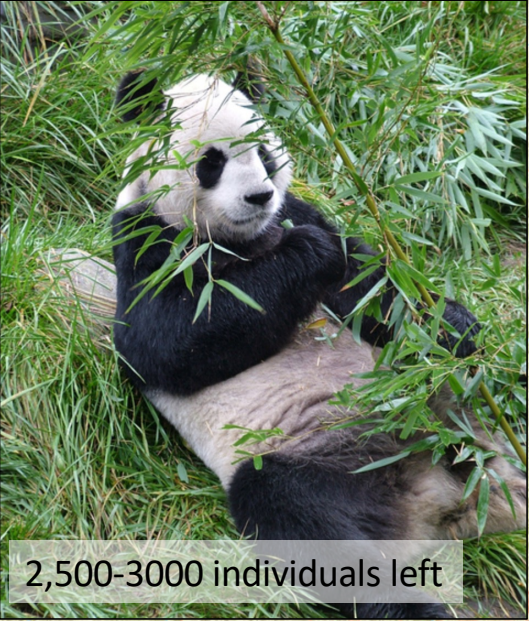
Boise State University

Spring 2022

WHY THIS STUDY?

- First study to conduct *de novo* genome assembly of an eukaryote based solely on NGS data.
- Study used an iterative genome assembly approach (implemented in SOAPdenovo) based on fragments' library sizes varying from 150 bp to 10 kb.
- Quality of assembly was assessed using BAC sequencing (same as Human Genome Project; see References tab on course website).
- This study also pioneered a protocol to assess species genome size and complexity based on NGS fragments (using k-mers; Chapter 4).

WHY THE GIANT PANDA?



2,500-3000 individuals left

A species on the edge of extinction

CHINA

Hanzhong

Ankang

Mianyang

Deyang

Chengdu

Nanchong

Daxian

Wanxian

Linshui

Kangding

Leshan

Neijiang

Zigong

Yibin

Wangda

Changde

Image Unavailable

Mammalia > Carnivora > Ursidae
Ailuropoda melanoleuca
Giant Panda
[Download Spatial data](#)

>Back to Red List Page

NE DD LC NT **< VU >** EN CR EW EX
VULNERABLE

Extant (resident)

Possibly Extant (resident)

BROWSE IMAGES

[ARKive \(0 found\)](#)

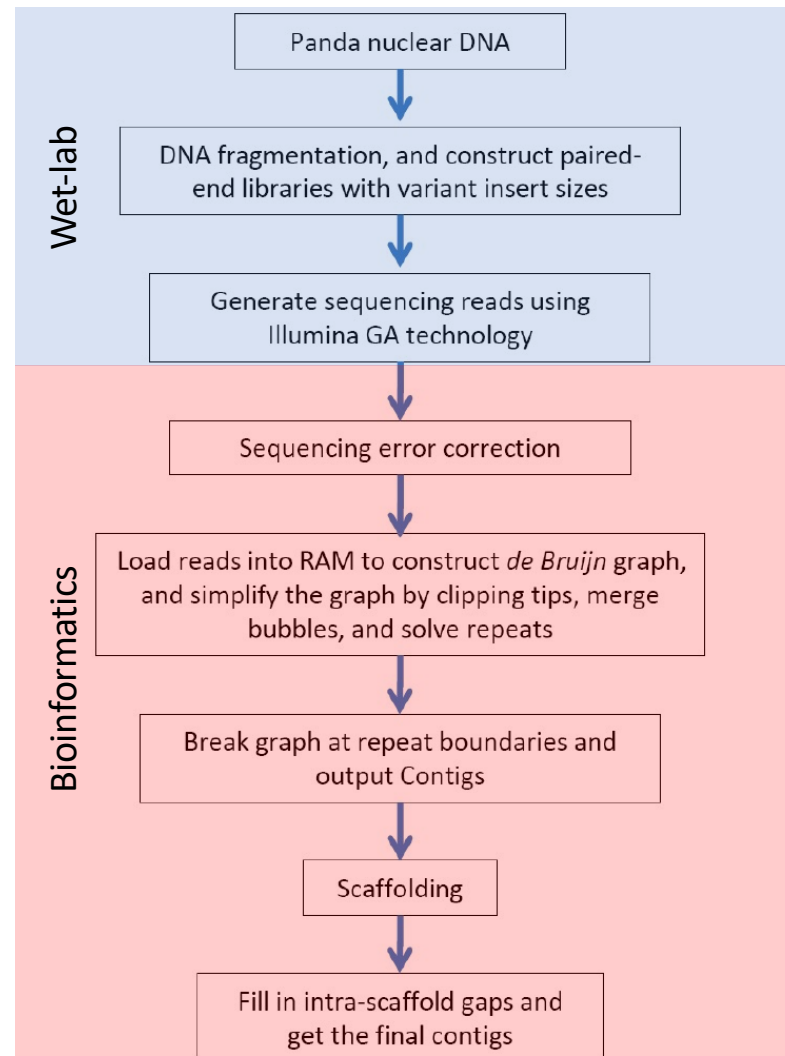
Dajun Wang 2016. Ailuropoda melanoleuca. The IUCN Red List of Threatened Species. Version 2017-3

With very unusual biological and behavioral traits:

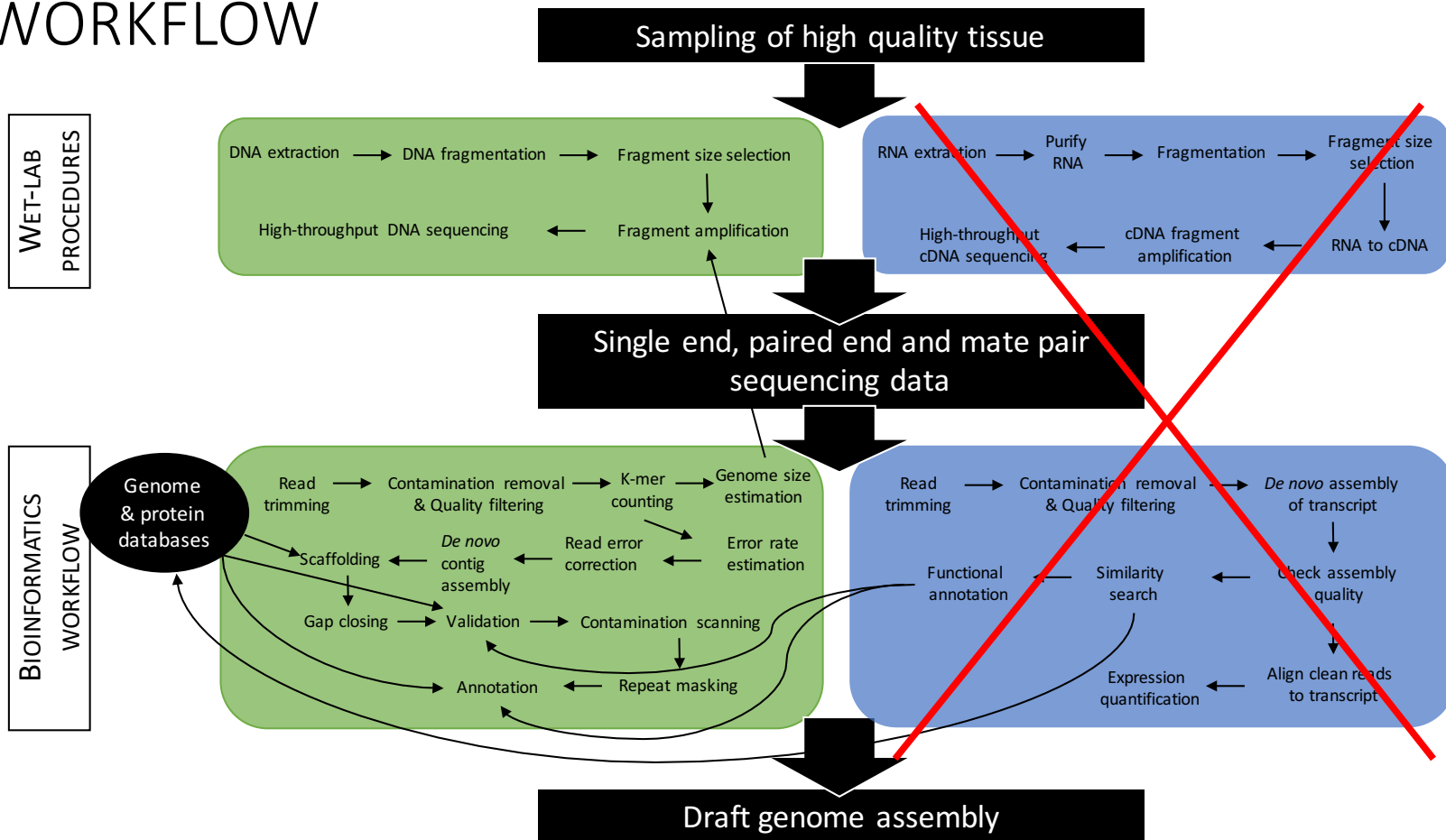
- Restricted diet (bamboo).
- Low fecundity rate.
- Controversy about phylogenetic position.
- $2n = 2x = 42$ (one pair of sex chromosomes).
- Authors selected a 3-year-old female from Chengdu breeding center (China).

Esri, HERE, Garmin, NGA, USGS, IUCN

WORKFLOW



WORKFLOW



RESULTS

- 37 PE sequencing libraries were constructed with insert sizes of: 150 bp, 500 bp, 2 kb, 5 kb and 10 kb.
- 176 Gbp of raw data generated (73-fold genome coverage).
- After cleaning, used 134 Gb for *de novo* assembly (56-fold genome coverage).
- Reads were mapped back onto assembly to assess accuracy and coverage (99% of bases had coverage of at least 20x). These data will later be used to call SNPs.
- Data freely available on NCBI.

RESULTS – ESTIMATING GENOME SIZE & COMPLEXITY

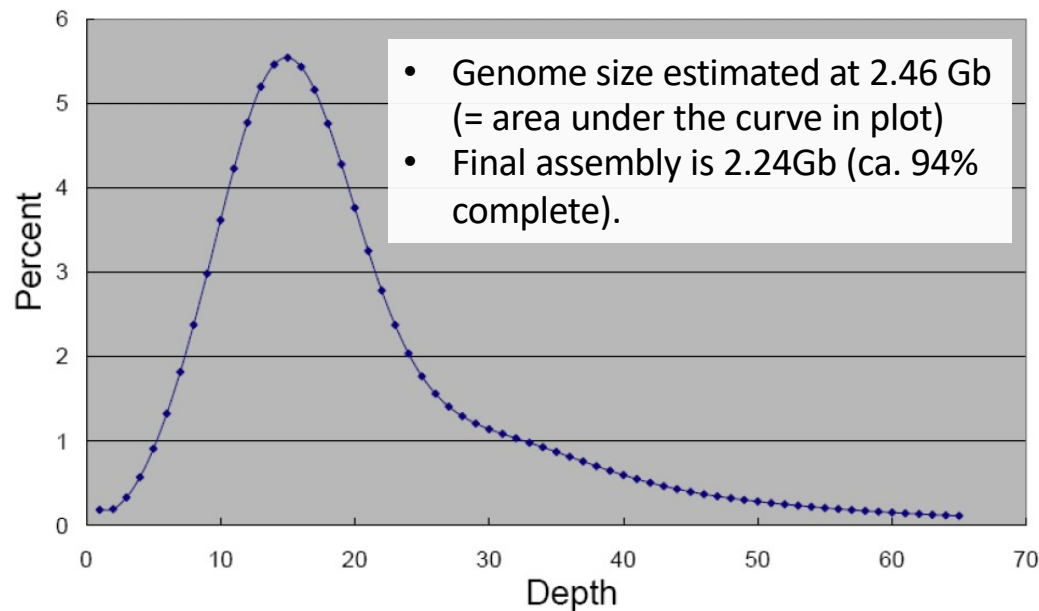


Figure S8 | Distribution of 17-mer frequency in the raw sequencing reads. We used all reads from the short insert-size libraries (<500bp). The peak depth is at 15X. The peak of 17-mer frequency (M) in reads is correlated with the real sequencing depth (N), read length (L), and kmer length (K), their 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.

MATERIAL & METHODS

- ***De novo* assembly:**

- ✓ Data QCs.

- ✓ Iterative *de novo* assembly approach (using SOAPdenovo):

1. Produce contigs based on small insert-size libraries (<500bp).
2. Used PE information, step by step from the shortest (150 bp) to the longest (10kb) insert size to join contigs into scaffolds.
3. Fill small gaps between contigs (mostly repetitive regions) by mapping reads onto assembly.

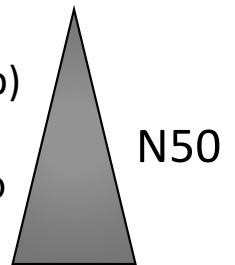


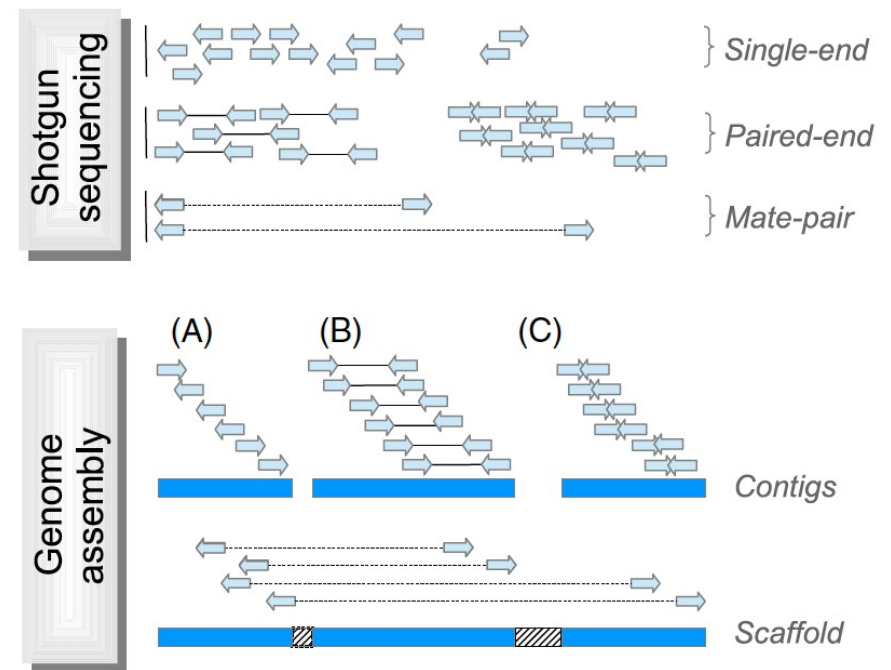
Table 1 | Summary of the panda genome sequencing and assembly

Step	Paired-end insert size (bp)*	Sequence coverage (×)†	Physical coverage (×)†	N50 (bp) ‡	N90 (bp) ‡	Total length (bp)
Initial contig				1,483	224	2,021,639,596
Scaffold 1	110–230; 380–570	38.5	96	32,648	7,780	2,213,848,409
Scaffold 2	Add 1,700–2,800	8.4	151	229,150	45,240	2,250,442,210
Scaffold 3	Add 3,700–7,500	6.5	450	581,933	127,336	2,297,100,301
Scaffold 4	Add 9,200–12,300	2.6	373	1,281,781	312,670	2,299,498,912
Final contig	All	56.0	1,070	39,886	9,848	2,245,302,481

N50 size of contigs or scaffolds was calculated by ordering all sequences then adding the lengths from longest to shortest until the summed length exceeded 50% of the total length of all sequences.

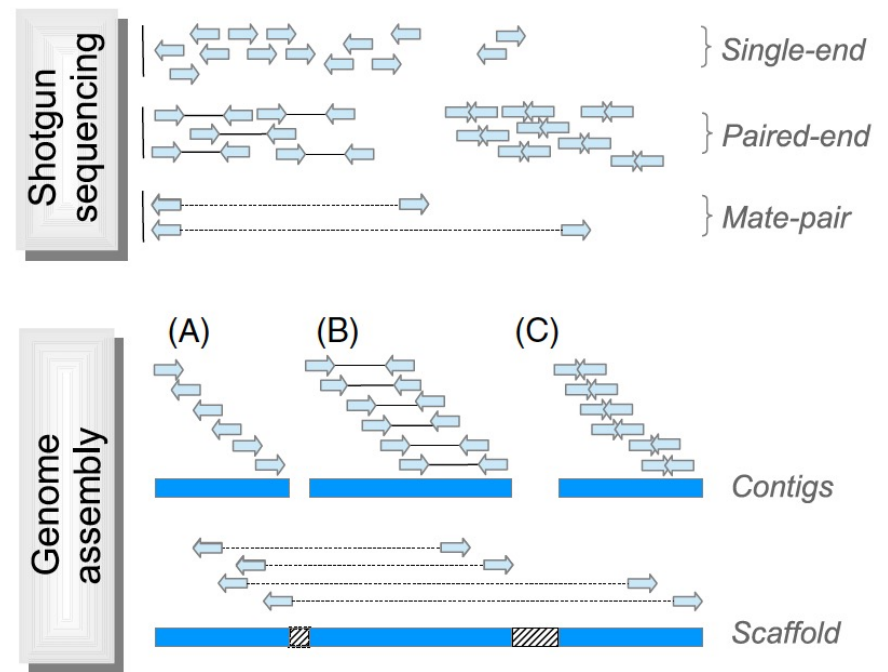
WHAT IS N50?

- N50 is a measure to describe the quality of assembled genomes that are fragmented in contigs/scaffolds of different length.
- **The N50 is defined as the minimum contig/scaffold length needed to cover 50% of the assembled genome.**



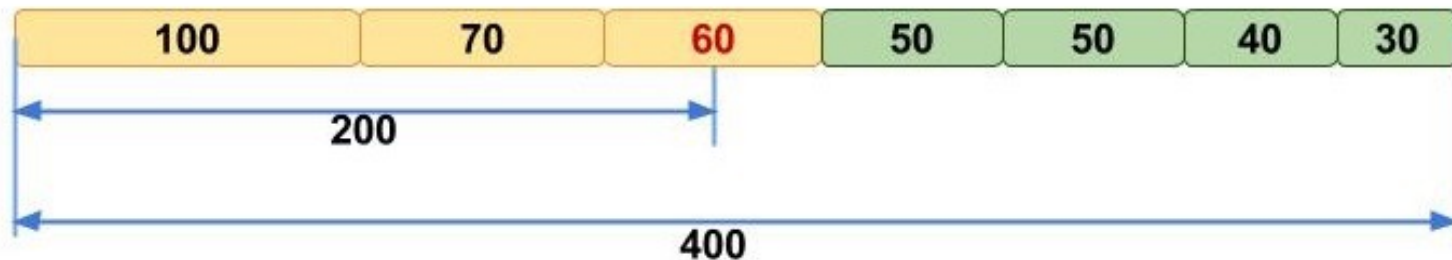
WHAT IS N50?

- **Half of the genome sequence is in contigs/scaffolds larger than or equal to the N50 contig size.**
- Or, that the sum of the lengths of all contigs of size N50 or longer contain at least 50 percent of the total genome sequence.



WHAT IS N50?

- **N50 size of contigs or scaffolds is calculated by:**
 - I. Ordering all sequences (based on length),
 - II. Summing the lengths of sequences (starting by the longest) until the sum exceeds 50% of the total length of all sequences.



Example with 7 contigs:

Total length contigs = 400 bp,

50% length contigs = 200 bp

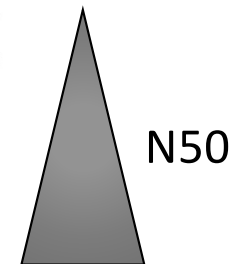
→ **N50 = 60bp**

RESULTS

Iterative approach improving assembly as shown by N50

Table 1 | Summary of the panda genome sequencing and assembly

Step	Paired-end insert size (bp)*	Sequence coverage (×)†	N50 (bp) ‡	N90 (bp) ‡	Total length (bp)
Initial contig			1,483	224	2,021,639,596
Scaffold 1	110–230; 380–570	38.5	32,648	7,780	2,213,848,409
Scaffold 2	Add 1,700–2,800	8.4	229,150	45,240	2,250,442,210
Scaffold 3	Add 3,700–7,500	6.5	581,933	127,336	2,297,100,301
Scaffold 4	Add 9,200–12,300	2.6	1,281,781	312,670	2,299,498,912
Final contig	All	56.0	39,886	9,848	2,245,302,481

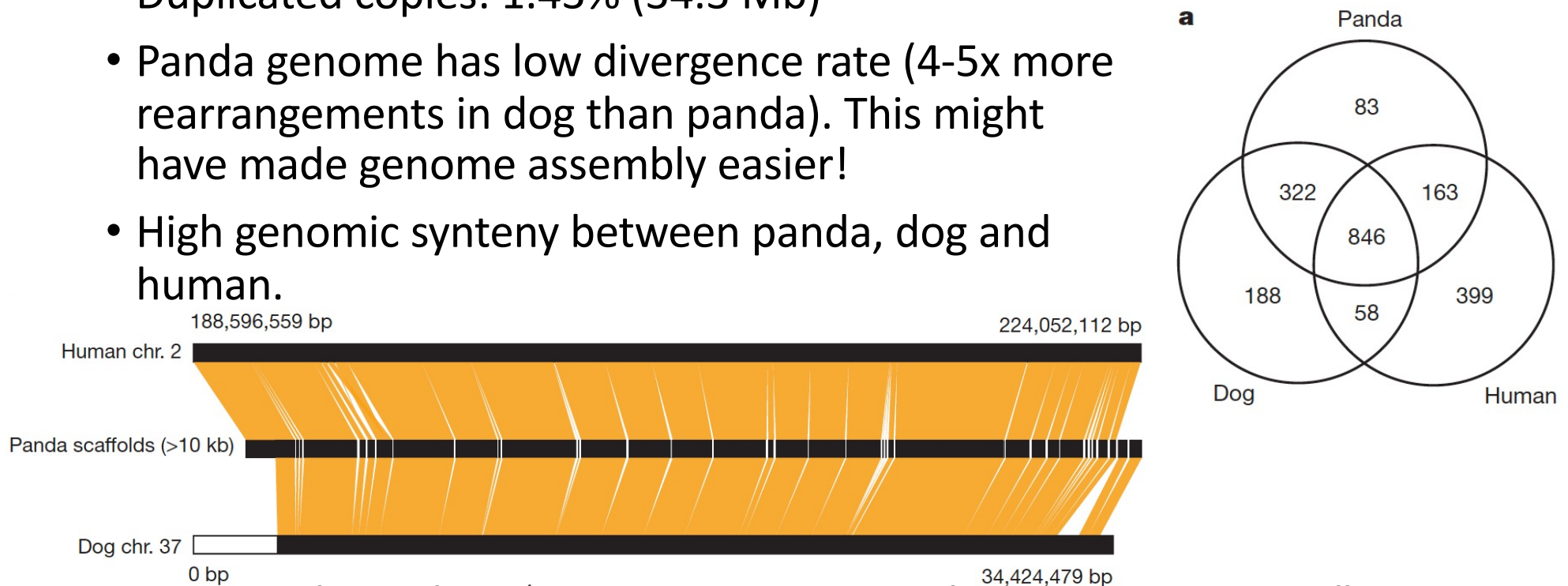


N50 size of contigs or scaffolds was calculated by ordering all sequences then adding the lengths from longest to shortest until the summed length exceeded 50% of the total length of all sequences.

FEATURES OF THE PANDA GENOME

- Transposable elements: 36%
- Duplicated copies: 1.43% (34.3 Mb)
- Panda genome has low divergence rate (4-5x more rearrangements in dog than panda). This might have made genome assembly easier!
- High genomic synteny between panda, dog and human.

Figure 2 | Conserved sequences among the panda, dog and human genomes. a, The total lengths of aligned and unaligned non-repetitive sequences. Each of the three genomes contains 1.4 Gb of non-repetitive sequences. Pairwise whole-genome alignment was performed using Blastz.



Synteny: conservation of blocks of genes/DNA between two or more sets of chromosomes belonging to different species.

FEATURES OF THE PANDA GENOME

- Gene predictions (based on human genome) made with Genscan & Augustus inferred respectively 44,428 and 29,238 gene loci.
- Dynamic evolution of orthologous gene clusters suggested greater amount of gene contraction than expansion.
- Genes that underwent most expansion are associated with receptor activity.

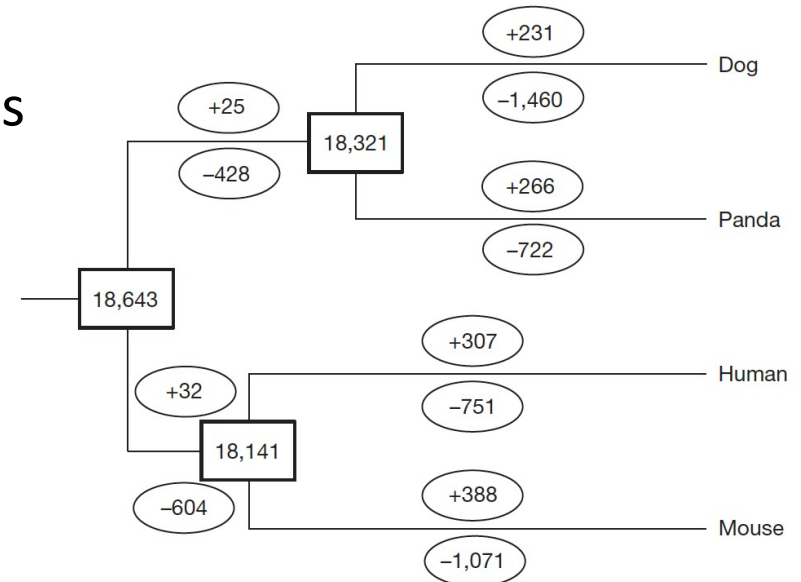


Figure 3 | Dynamic evolution of orthologous gene clusters. The estimated numbers of orthologue groups in the common ancestral species are shown on the internal nodes. The numbers of orthologous groups that expanded or contracted in each lineage after speciation are shown on the corresponding branch, with '+' referring to expansion and '-' referring to contraction.

Orthologous genes: Genes diverging after evolution gives rise to different species (i.e. they follow speciation events).

FEATURES OF THE PANDA GENOME

- Panda has all the genes to sustain a carnivorous diet.
- There are no genes involved in digesting leaf matter. → Panda's bamboo diet is not dictated by its genome, but most likely by its gut microbiome.
- Loss-of-function of the *T1R1* gene might have prevented the panda from expressing a functional umami taste receptor, which may partly explain why the panda is primarily herbivorous despite being part of a carnivorous lineage.

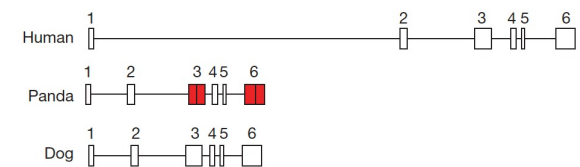
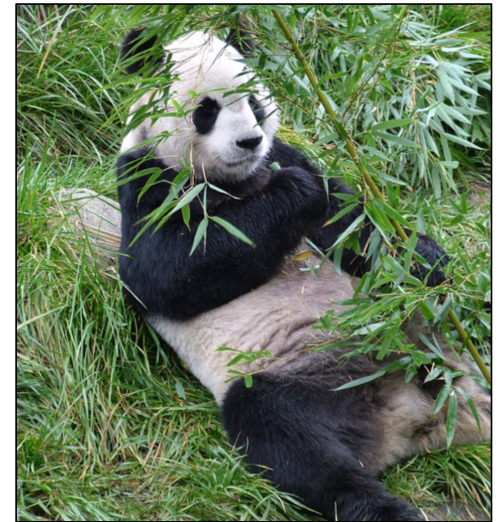


Figure 4 | Structure of the umami receptor *T1R1* gene. Two frameshift mutations occurred in the third and sixth exons (red) of the panda *T1R1* gene. The third exon contained a 2-bp ('GG') insertion; the sixth exon contained a 4-bp ('GTGT') deletion.

Umami, or savory taste: One of the five basic tastes, which has been described as brothy or meaty and is associated with glutamate.

FEATURES OF THE PANDA GENOME

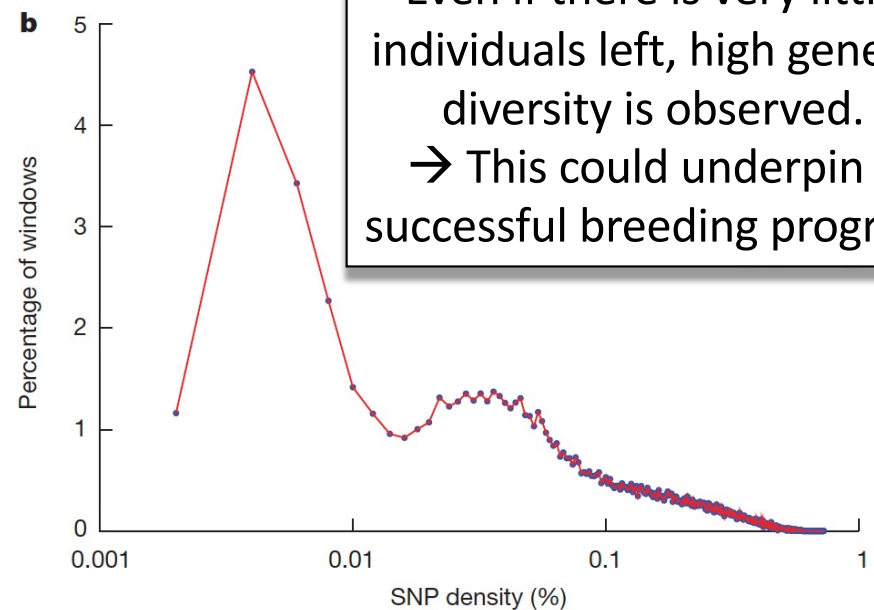
- 2.7Mio of heterozygous SNPs in Panda diploid genome.
- 1.95 times higher than the rate observed in the human genome.

a

	Analysed regions	Heterozygote no.	Heterozygote rate ($\times 10^{-3}$)
Genome	2,036,140,541	2,682,349	1.32
Autosomes	1,945,074,681	2,621,978	1.35
Chr. X	91,065,860	60,371	0.66
CDS	29,559,494	19,115	0.65
Autosomes	28,447,156	18,726	0.66
Chr. X	1,112,338	389	0.35

Figure 5 | Panda heterozygous SNP density. **a**, Statistics of identified heterozygous SNPs. Analysed regions are the genomic regions with proper unique read coverage that were used for heterozygote detection. The panda X-chromosome-derived scaffolds were identified by Blastz alignment to the dog X chromosome. **b**, Distribution of heterozygosity density in the panda diploid genome. Heterozygous SNPs between the two sets of chromosomes of the panda diploid genome were annotated, then non-overlapping 50-kb windows were chosen and heterozygosity density in each window was calculated.

b



Even if there is very little individuals left, high genetic diversity is observed.
→ This could underpin a successful breeding program

QUESTIONS — DISCUSSIONS IN BREAKOUT ROOMS

- **Q1:** Is there any evidence in the annotated genome sequence of the giant panda supporting its bamboo diet?
- **Q2:** What is the main outcome of the SNPs analysis conducted on the giant panda? What does this mean for the conservation of this threatened species?