



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Doctoral Dissertation

De novo Genome Reference Assembly

Hak-Min Kim

Department of Biomedical Engineering

Graduate School of UNIST

2019

De novo Genome Reference Assembly

Hak-Min Kim

Department of Biomedical Engineering

Graduate School of UNIST

De novo Genome Reference Assembly

A dissertation
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Hak-Min Kim

6. 11. 2019

Approved by



Advisor

Jong Bhak

De novo Genome Reference Assembly

Hak-Min Kim

This certifies that the dissertation of Hak-Min Kim is approved.

6. 11. 2019

signature



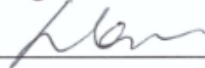
Advisor: Jong Bhak

signature



Cheol-Min Ghim: Thesis Committee Member #1

signature



Dougu Nam: Thesis Committee Member #2

signature



Semin Lee: Thesis Committee Member #3

signature



Seung woo Cho: Thesis Committee Member #4;

Abstract

Recent advancements in sequencing technologies have helped to sequence the complete genomes of many species encompassing all the kingdoms of life. However, the assembly of large and complex genomes remains challenging. Here, I report the first genome assemblies of Amur leopard (*Panthera pardus*) and Nomura's jellyfish (*Nemopilema nomurai*), which were processed by different strategies of sequencing platforms and downstream analysis methods. Genome survey results of the two species indicate that the leopard genome is much larger than that of the jellyfish but showed a relatively low heterozygosity. The leopard and jellyfish genomes were sequenced by the second- (Illumina short reads) and third-generation (PacBio SMRT long reads) sequencing technologies, respectively. Recent studies indicate that the sequencing platform has the most influence on determining the genome assembly quality and current sequencing technologies have clear limitations. Short-read based sequencing has a problem in resolving repeats, and long-read based sequencing is not suitable for large genomes because it requires a high sequencing coverage (>50X) due to high error rates. Therefore, I propose that a hybrid sequencing strategy is the most efficient method for reducing sequencing and computational cost. The difference in the evolutionary positions of the two species shows the necessity for different analytical approaches. I demonstrate that the leopard, which has an evolutionary distance of less than 10 million years from other Felidae species, could be subject close species comparative genomics (CSCG), such as homology-based comparative, positive selection, unique amino acid changes, and highly conserved region analyses, whereas the jellyfish genome was analyzed under distant species comparative genomics (DSCG), such as conserved protein domains and absence/presence of conserved genes, because the evolutionary distance to other cnidarian genomes was more than 200 million years. It clearly suggests that it is necessary to use radically different strategies depending on their evolutionary positions. Through a comparison between the two very different species, this study provides guidelines to determine the optimal strategies for a new genome reference assembly.

Contents

List of Figures	
List of Tables	
Nomenclature	
I . Introduction	1
II . Methods	5
2.1 Sample preparation	5
2.2 Genome sequencing and assembly	5
2.3 Genome annotation	7
2.4 Comparative evolution analyses	8
2.5 Genetic diversity and demographic history	10
2.6 Protein domain analyses	10
III. Results & Discussion	12
3.1 Genome survey and sequencing.....	12
3.2 Genome assembly	16
3.2.1 Leopard genome assembly	18
3.2.2 Jellyfish genome assembly	22
3.3 Genome annotation	27
3.4 Comparative genomics analysis of leopard and jellyfish.....	33
3.4.1 Comparative genomics analysis of leopard genome	34
3.4.2 Comparative genomics analysis of jellyfish genome	51
IV. Conclusions	61
References	63
Acknowledgements	75
Appendix	76

List of Figures

Figure 1. Distribution of K-mer frequency of leopard	13
Figure 2. Distribution of K-mer frequency of jellyfish	14
Figure 3. Species and sub-species identification for three leopard samples	18
Figure 4. Length distribution of PacBio SMRT reads	22
Figure 5. Schematic overview of the <i>Nemopilema nomurai</i> genome assembly process	23
Figure 6. Length distribution of Illumina TruSeq synthetic long reads	25
Figure 7. GC content distributions of leopard genome	28
Figure 8. GC content distributions among cnidarian genomes	32
Figure 9. Felidae-specific amino acid changes in DNA repair system	38
Figure 10. Felidae-specific amino acid change in MEP1A protein	39
Figure 11. Felidae-specific amino acid change in ACE2 protein	40
Figure 12. Felidae-specific amino acid change in PRCP protein	40
Figure 13. Highly conserved regions in Felidae, Hominidae, and Bovidae. Highly conserved regions in the same family species were identified by calculating the ratios between numbers of conserved and non-conserved positions	41
Figure 14. Genetic diversity in Felidae species	50
Figure 15. Expanded domains in <i>Nemopilema nomurai</i> based on Pfam domain annotation	51
Figure 16. Multiple sequence alignment of homeobox domains for Hox and ParaHox genes with human, fruit fly, and cnidarians	54
Figure 17. Phylogenetic analysis of Hox and ParaHox homeobox domains with human, fruit fly, and cnidarians	55
Figure 18. Presence and absence of Hox genes in cnidarians	56
Figure 19. Arrangements of Hox and ParaHox genes in cnidarians	56
Figure 20. Phylogenetic tree using Maximum likelihood of Wnt proteins	58
Figure 21. A primordial cluster of three <i>Wnt</i> gene (<i>Wnt1–Wnt6–Wnt10</i>) pattern of cnidarians	60

List of Tables

Table 1. First- to Fourth-generation of sequencing technologies	1
Table 2. Genome assembly quality standards	3
Table 3. Filtered Illumina sequence information of leopard	12
Table 4. Filtered Illumina sequence information of jellyfish	13
Table 5. Estimated genome size of leopard based on <i>K</i> -mer frequency	14
Table 6. Estimated genome size of jellyfish based on <i>K</i> -mer frequency	15
Table 7. Performance summary of short read assemblers	17
Table 8. The leopard genome assembly statistics	19
Table 9. Sample information of wild Amur leopards and Amur leopard cat used in this study	20
Table 10. Illumina TruSeq Synthetic Long Reads from two wild Amur leopard individuals	20
Table 11. Assembly results in five Felidae genomes	21
Table 12. Leopard assembly quality assessment using self-alignments	21
Table 13. Length distribution of PacBio SMRT reads	23
Table 14. Contig assembly statistics using PacBio SMRT reads	24
Table 15. Scaffold assembly statistics using PacBio SMRT reads and Illumina mate-pair reads	24
Table 16. Illumina TruSeq Synthetic Long Reads statistics	26
Table 17. Assembly quality assessment by mapping Illumina reads to <i>Nemopilema</i> assembly	26
Table 18. Assembly statistics of nine metazoans and choanoflagellate	26
Table 19. Statistics regarding predicted protein-coding genes in leopard genome	27
Table 20. Statistics regarding transposable elements (TEs) in leopard genome	28
Table 21. Assembly and annotation quality assessment of leopard genome using single-copy orthologs mapping approach	29
Table 22. Sequencing statistics regarding two wild Amur leopards and an Amur leopard cat	29
Table 23. Transcriptome sequence statistics of the jellyfish	30
Table 24. Statistics of post-filtered protein-coding gene properties in metazoans and holozoan	30
Table 25. Gene-set quality assessment of jellyfish using a single-copy ortholog mapping approach	31
Table 26. Repeat annotation of cnidarians	32

Table 27. GO enrichment of positively selected genes in leopard	34
Table 28. GO enrichment of shared positively selected genes in Felidae	35
Table 29. Variants statistics regarding mapping of Felidae raw reads to the cat reference (Felis_catus_8.0)	36
Table 30. GO enrichment of Felidae-specific genes having function altering amino acid changes .	37
Table 31. KEGG pathway enrichment of Felidae-specific genes having function altering amino acid changes	37
Table 32. Variants statistics regarding mapping of Felidae raw reads to the cat reference (Felis_catus_8.0)	42
Table 33. Variants statistics regarding mapping of Hominidae and Bovidae raw reads to the human and cow references	42
Table 34. Statistics regarding highly conserved regions in Felidae, Hominidae, and Bovidae genomes	43
Table 35. GO enrichment of shared genes in the highly conserved regions of Felidae, Hominidae, and Bovidae	44
Table 36. KEGG pathway enrichment of shared genes in the highly conserved regions of Felidae, Hominidae, and Bovidae	45
Table 37. GO enrichment of Felidae-specific genes in the highly conserved regions	46
Table 38. KEGG pathway enrichment of Felidae-specific genes in the highly conserved regions ..	47
Table 39. GO enrichment of Hominidae-specific genes in the highly conserved regions	47
Table 40. GO enrichment of Bovidae-specific genes in the highly conserved regions	48
Table 41. KEGG pathway enrichment of Bovidae-specific genes in the highly conserved regions .	48
Table 42. Presence of Hox, Hox-related, and ParaHox homeobox domains in Cnidaria	52
Table 43. Distribution of <i>Wnt</i> genes among cnidarians	59

Nomenclature

AAC, amino acid change
CSCG, close species comparative genomics
DSCG, distant species comparative genomics
Gb, giga bases
HCR, highly conserved region
KOREF, the Korean Reference genome
KPGP, Korean personal genome project
Mb, mega bases
NGS, next-generation sequencing
OLC, overlap-layout consensus
PacBio, Pacific Biosciences
PCR, polymerase chain reaction
PSG, positively selected gene
PSMC, pairwise sequentially Markovian coalescent
SMRT, single molecule real time sequencing technology
SNV, single nucleotide variation
TSLR, Illumina TruSeq synthetic long reads

I . Introduction

DNA sequencing technologies continue to advance and offer many utilities for biological researchers. The rapid improvement in sequencing technologies has helped to sequence many complete genomes, including the human genome, such as the Korean Reference genome (KOREF)¹ and tiger, whale, vulture, and red bat genomes in Korea²⁻⁵. Since Frederick Sanger invented the first chain termination sequencing method, there have been numerous next-generation sequencers, most of which produce only short sequences. Recently, however, very long-read based the third- and fourth-generation sequencing technologies⁶ have become cost-effective and useful (Table 1).

Table 1. The first- to fourth-generation sequencing technologies.

Generation	Technology	Released year	Method	Pros	Cons
First	ABI/Life technologies	1998	CE-Sanger	High accuracy. Good ability to call homopolymer and repeats regions	High cost of Sanger sample preparation; Low throughput;
Second	Roche 454	2005	Pyrosequencing.	Long reads. Useful for assemble the genome.	Relatively low throughput; difficulties in reading homopolymers.
Second	Illumina/ Solexa	2006	Sequencing by synthesis.	Very high throughput. Useful for gene expression analyses.	Short read length with relatively low quality in the read ends.
Second	ABI SOLID	2006	Sequencing by ligation.	Very high throughput. Low reagent cost.	Long sequencing time. Difficulty of data analysis.
Third	Ion torrent PGM	2010	Semiconductor sequencing.	Long reads. DNA synthesis reactions work under natural conditions.	Sequential cleaning steps can cause accumulation of errors. Difficult to read through repetitive and homopolymer regions.
Third	PacBio SMRT	2011	Single-molecule real time sequencing.	Very long average read lengths. No amplification of sequencing fragments.	High error rate.
Fourth	Nanopore	2015	Nanopore exonuclease sequencing.	Very long read. very fast.	High indel and base detection error rate.

The sequences generated by these improved technologies enable comparative analysis by aligning to pre-assembled genome sequence called reference genomes. However, species lacking a genome sequence can be analyzed in two ways. The first approach is to align to the reference genome of the closest species⁷. While this method is very efficient, the evolutionary distance must be close enough to the reference genome for accurate analysis. That is, if the two species are too distantly related then the analysis is limited to evolutionarily conserved genes. The second approach is to assemble the genome sequence *de novo*⁸. Sequence assembly, especially genome assembly, refers to ordering and merging short fragment of DNA sequences to reconstruct the original genome sequence. Although genome assembly is still challenging, it is essential for comparative analysis of various species.

The assembly quality is affected by various factors, such as heterozygosity, %GC content, segmental duplication, whole genome duplication, repeat composition, polyploidy, and sequencing bias. These genomic factors make genome assembly incomplete. The assembly quality is also affected by the incompleteness of analytic tools, such as assemblers, scaffolding and gap filling tools. Although their performance is steadily improving, it still affects the quality of assembly. After assembling the genome sequence, quality assessment is essential prior to use for downstream analysis. The method for assessing the quality of the assembled genome uses the number of sequences and size of each contig and the scaffold of the assembly, the total assembled size, and the N50, a weighted central statistical value that includes 50% of the entire assembly in a contig or scaffold that is equal to or greater than this value. In addition, the quality of the assembled genome sequence can be assessed using an ortholog set of genes⁹. This ortholog gene set is composed of single-copy genes that are evolutionarily well-conserved and have a relatively low selection pressure. Therefore, this set of genes allows us to assess the assembled genome that how many single copy orthologous genes are presented, duplicated, fragmented and missed. The genome assembly at the initial stage is called the "draft genome". There are several criteria and grades set by the Genomic Standards Consortium and Human Microbiome Jumpstart Reference^{10,11}, ranging from "standard draft" to "finished". Most of the first assemblies are either high-quality drafts or improved high-quality drafts (Table 2). At least an improved high-quality draft assembly is required for comparative genomics.

Table 2. Genome assembly quality standards

Grade	Description
Standard draft	Genome assembly with the minimum quality. It can be incomplete and contain unfiltered sequences derived from sequencing contaminants.
High-quality draft	Draft genome assembly with a coverage of at least 90% of the genome. In this assembly, effort has been made to exclude contaminating sequences, although it can contain sequence errors, misassemblies, and contigs with incorrect order and orientation.
Improved high-quality draft	This genome assembly has undergone automated and/or manual work. It consists of a reduced number of contigs and scaffolds. It can still contain some undetectable misassemblies, mainly in regions of repeat, low-quality and base errors. This standard is adequate for comparative genomics.
Annotation-directed improvement	In this genome assembly, finishing work is targeted to clearly defined areas identified by an automated annotation pipeline. Repetitive regions are not resolved completely, and the assembly contains several errors, with an N50 > 50 kb.
Noncontiguous finished	The assembly has been subject to automated and manual improvement, and closure approaches have been successful in almost all gaps, misassemblies, and low-quality regions.
Finished	This genome assembly is the so-called “gold standard”. All sequences are complete and have been reviewed. All misassemblies have been resolved properly, and repetitive sequences have been ordered and correctly assembled.

The *de novo* genome assembly is accomplished by sequencing and assembling the entire genome of a species whose whole genome sequences have not yet been identified. Therefore, the completion of the genome assembly has several advantages that comparisons with closely related species can reveal the biological and evolutionary meanings of the species through genomic information. The Genome 10K Consortium was established¹² in response to the increased importance attributed to genome assembly for its role in promoting species diversity and conservation. The project aims to address fundamental questions in disease and biology, to preserve genetic information, and to identify species the most genetically at risk for extinction. The Vertebrate Genomes Project (VGP), a part of the Genome 10K Consortium, also aims to generate error-free reference genome assemblies of all 66,000 extant vertebrate species. In Korea, the Korea Post-Genome Project, a large-scale government-sponsored project launched in 2014, is currently in the process of discovering life resources using genomic information of animals, plants, and marine animals.

Here, I report genome assemblies of Amur leopard (*Panthera pardus*) and Nomura's jellyfish (*Nemopilema nomurai*), which have distinct genomic features*. Through a comparison of the two species, I present guidelines for strategic considerations when sequencing and assembling a new genome. The leopard genome was sequenced by Illumina short read, a second-generation technology, and the jellyfish genome was sequenced by PacBio SMRT long read, the third-generation technology. The genome of the leopard and jellyfish that I report here is the first published genome of species that play a vital role in their ecosystem. Notably, the leopard is classified as an endangered species, and the leopard genome is expected to be an important resource for coping with the endangered species. This study also examined how the two species evolved to adapt to each environment through comparative genomic analyses.

*This doctoral dissertation is an addition based on the following papers that the author has already published.

Soonok Kim, Yun Sung Cho, Hak-Min Kim, Oksung Chung, Hyunho Kim, *et al.* Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly. *Genome Biology* **2016**, 17, 211.

Hak-Min Kim, Jessica A. Weber, Nayoung Lee, Seung Gu Park, Yun Sung Cho, *et al.* The genome of the giant Nomura's jellyfish sheds light on the early evolution of active predation. *BMC Biology* **2019**, 17, 28.

II. Methods

2.1 Sample preparation

For leopard, the leopard sample used for a genome assembly was acquired from the Daejeon O WORLD Zoo of Korea. We confirmed that the leopard sample was ~30% admixture with North-Chinese leopard from pedigree information. Phylogenetic analyses on mitochondria genes of *NADH5* and *CYTB* also verified that the leopard sample is a hybrid between Amur and North-Chinese leopards. The four other Amur leopards and one Amur leopard cat samples were acquired from Russia and Korea, respectively.

For jellyfish, the medusa from one *Nemopilema nomurai* individual was collected at Tongyong Marine Science Station, KIOST (34.7699 N, 128.3828 E) on Sep. 12, 2013. The surface water temperature was 24 °C. After transport to the laboratory, the medusa bell and tentacles were dissected; and the tissues were snap frozen in liquid nitrogen and stored at -75 °C. The polyps of *Sanderia malayensis* were provided by Aqua Planet Jeju Hanwha (Seogwipo, Korea). The polyps were fed daily with freshly hatched *Artemia nauplii* in the animal culture room, which was maintained at 24±1 °C. The metamorphosed ephyrae in the summer season were fed with *Aurelia* sp.1. For DNA extraction, *Nemopilema* tissues were mortar-pulverized in liquid nitrogen and the powder was homogenized in a lysis solution [2% CTAB, 1.4 M NaCl, 100 mM Tris-Cl (pH 8.0), 20 mM EDTA, 1% β-mercaptoethanol], and incubated at 65°C for 1 h. The same volume of a phenol:chloroform:isoamylalcohol (23:24:1) mixture was added to denature the proteins and the phases were separated by centrifugation at 12,000 rpm for 15 min at room temperature. The aqueous phase was saved and incubated at 37°C for 1 h after RNase A (30 mg/ml) was added. The DNA was extracted with a phenol:chloroform:isoamyl alcohol (25:24:1) mixture, a chloroform:isoamyl alcohol (24:1) mixture was added, and the samples were centrifuged at 12,000 rpm for 15 min at room temperature. A 1/10 volume of 3 M sodium acetate (pH 5.2) and the same volume of 100% ethanol were added into the retained aqueous phase. The precipitated DNA was washed using 70% ethanol and re-suspended in an appropriate volume of ion-exchanged ultrapure water. The DNA quantity was verified by the picogreen method using Victor 3 fluorometry, and agarose gel electrophoresis.

2.2 Genome sequencing and assembly

For leopard, we constructed 21 DNA libraries with different insert sizes (170bp, 400bp, 500bp, 700bp, 2 Kb, 5 Kb, 10 Kb, 15 Kb, and 20 Kb) according to the Illumina sample preparation protocol. The libraries were sequenced by Illumina HiSeq sequencers. HiSeq2500 for short insert libraries and

HiSeq2000 for long-mate pair libraries were used. I applied filtering criteria (PCR duplicated, adaptor contaminated, and $<Q20$ quality) to reduce the effects of sequencing errors. All filtered reads were corrected by K -mer analysis ($K=21$) and were used to assemble the genome using SOAPdenovo2¹³. The short insert size libraries (<1 Kb) were assembled into distinct contigs based on the various K -mer sizes ($K=27, 37, 43, 47, 53, 57, 63, 67, 73, \text{ and } 77$). Read pairs from all the libraries then were used to concatenate the contigs into scaffolds step by step from short to long insert size libraries. I closed the gaps using short insert size reads in two iterations. Only scaffolds exceeding 200 bp were used in this step. To reduce erroneous gap regions in the scaffolds, I aligned the $\sim 0.8\times$ Illumina TruSeq synthetic long reads (TSLRs) from two other wild Amur leopard individuals to the scaffolds using BWA-MEM¹⁴ and corrected the gaps with the synthetic long reads using in-house scripts.

For jellyfish, the jellyfish genome was sequenced using the followings: Pacific Biosciences (PacBio) single molecule real time sequencing (SMRT) reads, Illumina TruSeq synthetic long reads, and Illumina mate-pair reads. First, the extracted genomic DNA was sequenced to a $179\times$ average sequencing depth of coverage using a Pacific Biosciences RSII instrument with SMRT cell 8Pac V3 and DNA Polymerase Binding Kit P6 reagents (30 SMRT cells), as a major sequencing data source for a contig assembly. Additionally, a set of Illumina long mate-pair libraries (5 Kb, 10 Kb, 15 Kb, and 20 Kb) was generated. Sequencing and junction adaptor contaminated, PCR duplicated, and low quality ($<Q20$) reads were filtered out, leaving only highly accurate reads for genome assembly. Then, short insert size and long insert size reads were trimmed into 90 bp and 50 bp, respectively, to remove low quality end sequences. Also, 1.92 Gb ($\sim 9\times$ coverage) of Illumina TSLRs was generated to correct erroneous sequences in the PacBio long-read assembly and to close gap regions. Quality filtered PacBio long reads were assembled into contig sequences using the FALCON assembler¹⁵ with various read length cutoffs. To construct scaffold from contigs, I aligned the Illumina long-insert size libraries (5 Kb, 10 Kb, 15 Kb, and 20 Kb) to contig sets and constructed the scaffolds using SSPACE¹⁶. Gaps were filled by mapping the Illumina short-insert size reads by GapCloser¹³. I aligned TSLRs to scaffolds to fill the gaps and to correct erroneous sequences.

Two general approaches were applied to evaluate the quality of the assembled genome. First, a comparative matrix was constructed using general statistical values of the assembled genome and compared with the other species genomes. The values used in the comparison were such as the total size of the assembled genome, the number of sequences, the ratio of the gap, and N50. The second is quantitative measures for the evaluation of genome assembly based on single-copy orthologous genes from OrthoDB⁹.

2.3 Genome annotation

For leopard, the leopard genome was annotated for protein coding genes and repetitive elements. For the annotation of repetitive elements, I scanned the leopard genome for tandem repeats and transposable elements using Tandem Repeats Finder¹⁷, Repbase¹⁸, RepeatMasker¹⁹, and RepeatModeler²⁰. For the annotation of protein coding genes, *de novo* and homology-based gene prediction were conducted. For the homology-based gene prediction, I aligned cat, tiger, human, mouse, and dog protein sequences to leopard genome using TblastN²¹ with an E-value cutoff of 1E-5. The aligned sequences were clustered using GenBlastA²² and filtered by identity and coverage of >40% criterion. I used Exonerate software²³ to predict the gene structures. For the *de novo* gene annotation, AUGUSTUS software²⁴ was used. I filtered out possible pseudogenes (harboring premature stop-codons), genes shorter than 50-amino acids, and single exon genes that were likely to be derived from retro-transposition.

For jellyfish, I applied both *de novo* and empirical (homology- and evidence-based) gene prediction methods. For the homology gene prediction, I searched for sea anemone, hydra, sponge, human, mouse, and fruit fly protein sequences from NCBI database, and Cnidaria protein sequences from NCBI Entrez protein database using TblastN²¹ with an E-value cutoff of 1E-5. The aligned sequences were clustered using GenBlastA²² and filtered by coverage and identity of >40% criterion. I used Exonerate software²³ to predict the gene structures and exon hints were extracted using the `exonerate2hints.pl` script of the AUGUSTUS program²⁴. For the evidence-based gene prediction, I aligned the bell and tentacle RNA-seq reads to the repeat masked jellyfish genome assembly using the TopHat program²⁵. To remove redundantly aligned reads, I filtered the alignment results with the `--uniq` option using the `filterBam` command of AUGUSTUS. Intron hints were generated using the `bam2hints` command of AUGUSTUS. Protein-coding genes of jellyfish were determined using AUGUSTUS with the exon and intron hints with ≥ 30 amino acids criteria. Finally, I filtered the protein-coding genes that had breaks in the three-letter codon frame, premature stop codons, and ambiguous bases in the CDS. The completeness of genome assembly and gene annotation were evaluated by the commonly used single-copy orthologous gene mapping approach.

To annotate the repetitive elements in the assembled genomes, I scanned the genome for tandem repeats using the Tandem Repeats Finder database¹⁷. Transposable elements (TEs) were identified using both *ab initio*-based and homology-based approaches. The Repbase¹⁸ database version 19.03 was used for the homology-based approach to identify repeats using RepeatMasker¹⁹ and RMBlast. For the *ab initio*-based approach, I used RepeatModeler²⁰.

2.4 Comparative evolution analyses

For leopard, I constructed orthologous gene families for 17 mammalian genomes using OrthoMCL software²⁶. Genome sequences and protein-coding genes of human, mouse, cat, tiger, pig, cow, dog, horse, elephant, rabbit, giant panda, polar bear, killer whale, and opossum were obtained from the NCBI database. To calculate divergence time of the related species among mammals, I used four-fold degenerate sites of the single copy gene families using RelTime-CC²⁷ with the phylogenetic tree topology of published previous studies. The date of the node between human-dog was constrained to 97.5 million years ago (MYA) and cat-dog was constrained to 55 MYA according to divergence times from TimeTree database²⁸. A gene family contraction and expansion analysis was conducted using the CAFÉ program²⁹. I used the $P = 0.05$ criterion for significantly changed gene families.

To generate multiple sequence alignment among ortholog genes, PRANK³⁰ program was used. To estimate the dN/dS ratio (ω)³¹, the PAML package was used. The one-ratio model, which allows only a single dN/dS ratio for all branches, was used to estimate the general selective pressure acting among all species. A free-ratios model was used to analyze the dN/dS ratio along each branch. To further examine potential positive selection, the branch-site test of positive selection was conducted³². Statistical significance was assessed using LRTs with a conservative 10% FDR criterion³³. When I identified shared positively selected genes (PSGs), genomes in the same diet group (carnivores, omnivores, and herbivores) were excluded from background species; for example, I excluded other carnivore species from the background species, when I identified PSGs of leopard.

Also, I identified species-specific amino acid changes. To filter out biases derived from individual-specific variants, I used all of Felidae whole genome sequence data by mapping to the cat reference genome. The mapping was conducted using BWA-MEM, and variants were called using SAMTools program³⁴ with the “-d 5 -D 200” options. Function altering amino acid changes were predicted using PolyPhen-2³⁵ and PROVEAN³⁶ with the default options. Human protein sequences were used as templates in this step. A convergent amino acid change was defined, if all of target species has a same amino acid in same sequence position. The herbivore- or carnivore- specific function altered genes were identified, if all of target species has at least one function altering amino acid change in any position and all of different diet groups (carnivores or herbivores) has no function altering amino acid change. For functional enrichment tests, I used DAVID bioinformatics resources by using human genes as a background³⁷.

To characterize genetic variation in the genomes of three mammalian families (Felidae, Bovidae, and Hominidae), I scanned genomic regions that showed significantly reduced genetic variation by comparing variations of each window and whole genome (autosomes only). The Bovidae

and Hominidae genomes were obtained from the NCBI database and were mapped to cow (Bos_taurus_UMD_3.1.1) and human (GRCh38) references, respectively. Variants (SNVs and indels) were called using SAMtools. The numbers of heterozygous and homozygous positions within each 100 Kb window (bin size=100 Kb, step size=10 Kb) were estimated by calculating the numbers of conserved or non-conserved bases in the same family genomes. I only used windows that were covered more than 80 % of window size by all the mapped genomes. *P*-values were calculated by performing Fisher's exact test to test whether the ratio of homozygous to heterozygous positions in each window was significantly different from that of chromosomes. *P*-values were corrected using the Benjamini-Hochberg method³⁸, and only adjusted *P*-values of <0.0001 were considered significant. Only the middle 10 Kb of each significantly different window were considered as highly conserved regions. For functional enrichment tests of candidate genes by all the comparative analyses, I used the DAVID bioinformatics resources³⁷.

For jellyfish, orthologous gene clustering of protein-coding genes from eleven metazoans (*Nemopilema nomurai*, *Hydra magnipapillata*, *Nematostella vectensis*, *Acropora digitifera*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Danio rerio*, *Homo sapiens*, *Trichoplax adhaerens*, *Amphimedon queenslandica*, and *Mnemiopsis leidyi*) and one unicellular holozoan (*Monosiga brevicollis*: as an out-group) was conducted using the OrthoMCL (version 2.0.9) program²⁶. I found 306 single-copy gene families in the 12 species. To infer the jellyfish phylogeny, I used protein sequences of 100 single-copy gene families, and the PROTCATLG model in the RAxML (version 8.2.8) program³⁹. I also estimated the divergence time using the MCMCtree program with the approximate likelihood algorithm by PAML package³¹. The divergence date of the zebrafish-human node was constrained to 435 million years ago (MYA) and the fruit fly-roundworm was constrained to 743 MYA based on the TimeTree database²⁸. As expected, *Nemopilema* and *Hydra* formed a monophyletic clade that branched off the cnidarian stem before the common ancestor of Anthozoa arose. A gene contraction and expansion analysis was conducted using the CAFÉ program²⁹.

2.5 Genetic diversity and demographic history

For leopard, the Bovidae and Hominidae genome sequences were obtained from the NCBI database, and were mapped to cow (*Bos_taurus_UMD_3.1.1*) and human (GRCh38) references, respectively. Homozygous and heterozygous SNVs were called using SAMTools. Homozygous substitution rates were calculated by dividing the number of homozygous SNVs by corresponding species genome size (bp) and divergence time (MYA) from TimeTree database. Heterozygous SNV rates were calculated by dividing the number of heterozygous SNVs by the reference genome size.

To analyze the demographic histories of Felidae, I used the PSMC program⁴⁰. First, I extracted diploid genome sequence information from BAM files of seven big cats (three leopard, one tiger, one lion, one cheetah, and one snow leopard) and one small cat (leopard cat) re-sequencing data aligned to *Felis_catus_8.0*. To use only autosomal regions, I removed the read data aligned to sex chromosomes and mitochondrial genomes. I used PSMC options of `-N25 -t15 -r5 -p "4+25*2+4+6"` which have been previously used for great apes population history⁴¹. Generation times and mutation rates (per site, per year) were collected from previous studies^{2,42}.

2.6 Protein domain analyses

For jellyfish, I identified the homeobox domain regions in *Nemopilema* using the InterProScan program⁴³. The domain regions were predicted from the protein sequences using the InterProScan program with ProDom, Hamap, SMART, SUPERFAMILY, PRINTS, PANTHER, Gene3D, PIRSF, Pfam, ProSiteProfiles, TIGRFAM, ProSitePatterns, and Coils databases. To identify protein domains that are specifically expanded in the *Nemopilema* lineage, I conducted Fisher's exact test for Pfam categories comparing in-group counts (*Nemopilema*) to average counts in the outgroups (all other species in the analysis). This test was iterated over all domains, and the *P*-values obtained were corrected with a 5% false discovery rate (FDR) to identify the significantly expanded domains in *Nemopilema*. To visualize these expanded domains, counts were normalized by Z-score (row) and significantly expanded domains were plotted using the heatmap function in R. ParaHox and Hox genes were identified in *Nemopilema* by aligning the homeobox domain sequences of fruit fly and human to the identified *Nemopilema* homeobox domains. I considered only domains that were aligned to both the human and fruit fly. I also used this process for *Hydra*, *Nematostella*, and *Acropora* for comparison. Additionally, I added two Hox genes for *Hydra* and one Hox gene for *Acropora*, which are absent in NCBI gene sets, though they were present in previous study^{44,45}. ParaHox and Hox genes of *Clytia hemisphaerica*, a hydrozoan species with a medusa stage, were also added based on a previous study⁴⁶. Finally, a multiple sequence alignment of homeobox domains was conducted using MUSCLE, and a FastTree⁴⁷ maximum-likelihood phylogenetic tree was generated using the

PROTGAMMAJTT model.

Wnt genes of *Nematostella* and *Hydra* were obtained from previous studies^{48,49}, and those of *Acropora* were downloaded from the NCBI database. *Wnt* genes in *Nemopilema* were identified by searching for "wnt family" domain using the Pfam database. A multiple sequence alignment of *Wnt* genes was conducted using MUSCLE, and aligned sequences were trimmed using the trimAl program⁵⁰ with "gappyout" option. A phylogenetic tree was generated using RAxML with the PROTGAMMAJTT model and 100 bootstraps.

III. Results & Discussion

3.1 Genome survey and sequencing

Prior to assembling the genome, genome survey was first performed by *K*-mer analysis⁵¹. Genome surveys are used to understand the complexity of the genome and to predict its size. Both the leopard and jellyfish produced Illumina short reads, which were used for *K*-mer analysis (Tables 3 and 4). The leopard genome had one *K*-mer peak, while the jellyfish showed two distinct *K*-mer peaks (Figs. 1 and 2). The double peaks in the *K*-mer graph mean that the heterozygosity is high. In the case of jellyfish, the height of the first peak is similar to that of the second peak and thus shows a very high heterozygosity. It is reported that marine organisms typically show a high level of genome heterozygosity⁵².

Table 3. Filtered Illumina sequence information of leopard

Library		Number of remained read pairs	Trimmed read length	Remained total bases (bp)	Remained sequence depth (×)
170bp	L1	324,819,579	90	58,467,524,220	24.4
	L2	322,720,798	90	58,089,743,640	24.2
400bp	L1	463,815,627	90	83,486,812,860	34.8
500bp	L1	177,877,901	90	32,018,022,180	13.3
700bp	L1	247,339,040	90	44,521,027,200	18.6
	L2	233,469,831	90	42,024,569,580	17.5
2kb	L1	70,512,242	50	7,051,224,200	2.9
	L2	78,840,634	50	7,884,063,400	3.3
	L3	82,556,740	50	8,255,674,000	3.4
5kb	L1	46,062,964	50	4,606,296,400	1.9
	L2	55,322,387	50	5,532,238,700	2.3
	L3	55,745,264	50	5,574,526,400	2.3
10kb	L1	44,225,626	50	4,422,562,600	1.8
	L2	35,628,557	50	3,562,855,700	1.5
	L3	38,425,313	50	3,842,531,300	1.6
15kb	L1	25,137,484	50	2,513,748,400	1.0
	L2	23,451,001	50	2,345,100,100	1.0
	L3	9,374,114	51	956,159,628	0.4
	L4	6,094,053	51	621,593,406	0.3
20kb	L1	23,636,971	50	2,363,697,100	1.0
	L2	24,209,031	50	2,420,903,100	1.0
Total	-	2,389,265,157	-	158.6	

Table 4. Filtered Illumina sequence information of jellyfish

Insert-size	Library	Total number of reads	Read length (bp)	Total bases (bp)	Depth (\times , divided by 213Mb)	Total depth (\times)
400bp	L1_1	86,434,438	90	7,779,099,420	40.58	81.16
	L1_2	86,434,438	90	7,779,099,420	40.58	
5Kb	L1_1	21,407,082	50	1,070,354,100	10.05	20.10
	L1_2	21,407,082	50	1,070,354,100	10.05	
10Kb	L1_1	16,094,130	50	804,706,500	7.56	15.11
	L1_2	16,094,130	50	804,706,500	7.56	
15Kb	L1_1	9,090,529	50	454,526,450	4.27	8.54
	L1_2	9,090,529	50	454,526,450	4.27	
20Kb	L1_1	9,965,208	50	498,260,400	4.68	9.36
	L1_2	9,965,208	50	498,260,400	4.68	
Total	-	285,982,774	-	21,213,893,740		134.3

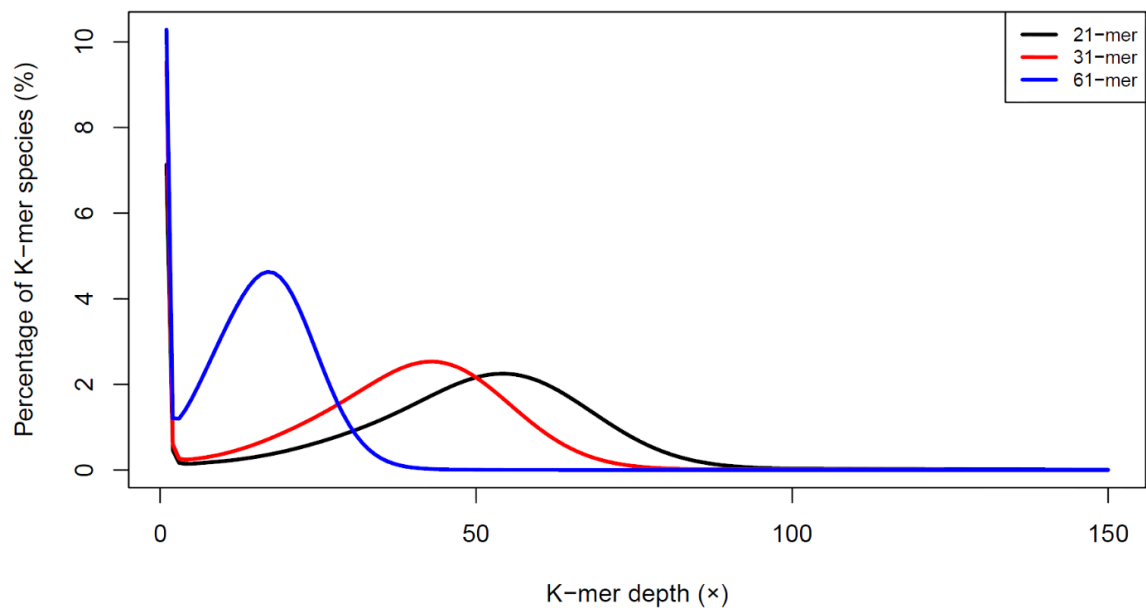


Figure 1. Distribution of K-mer frequency of leopard. The x-axis represents *K*-mer depth, and the y-axis represents proportion of *K*-mer species.

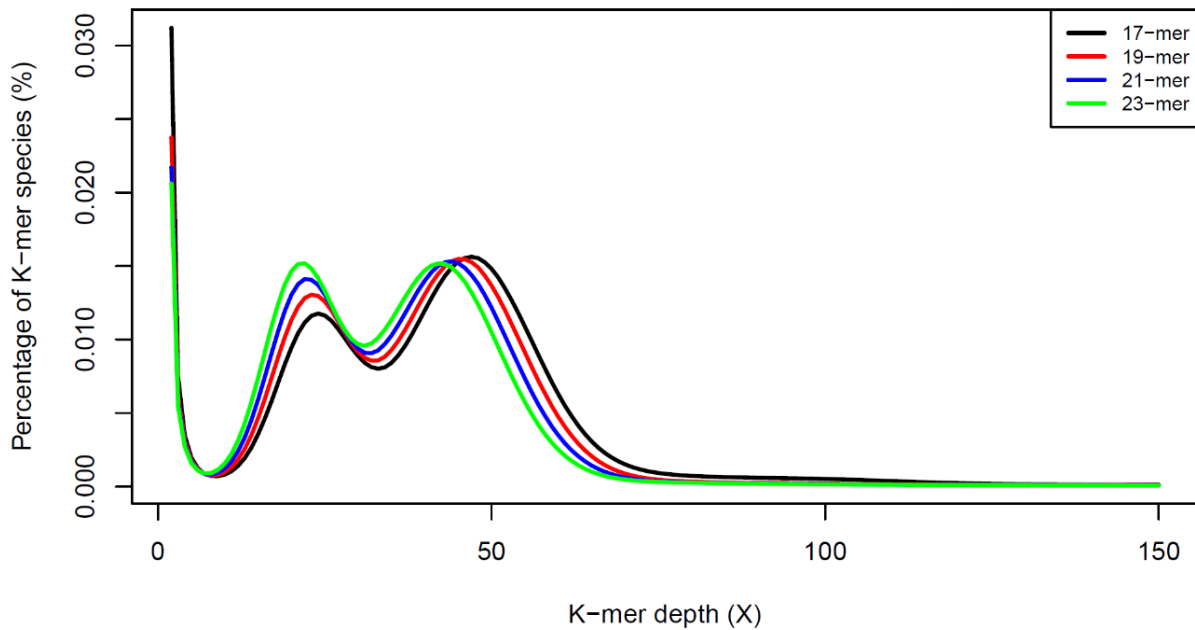


Figure 2. Distribution of K-mer frequency of jellyfish. The x-axis represents K -mer depth, and the y-axis represents proportion of K -mer species.

The genome size of the leopard was estimated at approximately 2.45 Gb (Table 5). This is similar to the genome size of tiger and lion, which belonging to the big cats. The jellyfish was estimated at 220 Mb (Table 6) and showed the smallest genome size when compared to the other cnidarians such as anemone, hydra, and coral⁵³⁻⁵⁵. Given the size and complexity of the genome, the leopard was determined to be sequenced with a short read based technology that could produce large quantities at relatively low cost. On the other hand, the jellyfish genome was determined to be sequenced by long read based technology due to its high level of heterozygosity.

Table 5. Estimated genome size of leopard based on K -mer frequency

K -mer size	Total K -mer count	Peak depth	Estimated genome size
21	116,054,812,460	54	2,149,163,194
31	95,405,247,995	43	2,218,726,698
61	41,733,211,831	17	2,454,894,814

Table 6. Estimated genome size of jellyfish based on *K*-mer frequency

<i>K</i> -mer size	Total <i>K</i> -mer count	Peak depth	Estimated genome size
17	9,934,145,023	47	211,364,788
19	9,732,013,270	45	216,266,962
21	9,520,986,025	44	216,386,046
23	9,302,429,653	42	221,486,420

A sequencing strategy for genome assembly should be established based on the results of the genome survey. Because current sequencing techniques have clear pros and cons, it is important to choose the optimal sequencing strategy based on the characteristics of the genome. First, short-read based sequencing technology is capable of producing a large amount of data and has a relatively low cost (see Table 1). Also, we can try many kinds of assemblers which have been improved for a long time. However, due to the short sequence length, this may be an inappropriate choice for the genomes with high repeat ratio or high heterozygosity. In addition, short-read sequencing technology poses a number of experimental problems, such as cloning, extreme GC bias, polymerase chain reaction (PCR), and sequencing errors⁵⁶.

Long-read based sequencing technology is safe from sequencing biases caused by high GC contents because the sequence length is very long, and there are no PCR steps. However, because it requires at least 50X of long-read data to produce a high-quality genome, long-read assemblies demand large sequencing and computational costs⁵⁷. Recent genome assembly projects using strictly long-read sequencing approaches have thus been applied to species with small genomes, such as viruses or bacteria⁵⁸⁻⁶⁰. The recent release of Oxford nanopore technology has made long-read based sequencing with affordable cost⁶¹, but it still shows a high error rate than PacBio SMRT sequencing.

3.2 Genome assembly

The basic strategy for *de novo* genome assembly for short-reads comprises three steps: i) contig assembly, ii) scaffolding and iii) gap filling. In the contig assembly step, the short reads are assembled as long consensus sequences (called contigs) without gaps and ambiguous bases. Then, in the scaffolding step, the contigs are connected by mate-pair sequences. The ordered set of connected contigs is called as a ‘scaffold’. Once the contigs are scaffolded, if there is no overlap between the contigs then spaces called ‘gaps’ remain between them, and unknown bases and approximate distances are estimated from the insert-size of the mate-pair reads. The gaps between the contigs are filled by short-reads or long-reads to complete the gap regions. The gap-filling step can be performed iteratively to improve the quality of the assembly.

Genome assembly algorithms can be divided into two types: graph method and greedy algorithm^{62,63}. A typical algorithm for the graph method is the de Bruijn graph method, which has been used for the short-read assembly by converting reads to K -mers. The graph using the K -mers can be simplified and significantly reduces the searching time for the optimal path^{62,64}. Assemblers using the de Bruijn graph method are SOAPdenovo2, SPAdes, and ALLPATHS-LG^{13,65,66}.

The overlap-layout consensus (OLC) algorithm, a typical method using the greedy algorithm, finds overlap between all reads, uses it to determine a layout of the reads, and then produces a consensus sequence. The OLC algorithm has been used for the long-read assembly. The newbler⁶⁷ and Celera assembler⁶⁸ both use the OLC algorithm. Assemblers, such as HGAP and Falcon, explicitly target the PacBio SMRT sequencing technology¹⁵. Canu and Hinge assemblers are developed for third- and fourth-generation sequencing technologies^{69,70}.

It is an important step to choose assembler for the *de novo* genome assembly. The first option is the sequencing platform. As mentioned above, the selection of assemblers is limited by the type of raw data (short read or long read). The second option is to select the appropriate assembler according to the characteristics of the genome. For example, high heterozygous genomes with short-read data should be paired with an assembler that addresses the heterozygous regions, such as Platanus⁷¹. Computing resources also need to be considered. The *de novo* assembly requires significant computing memory, storage, and long calculation times. In some cases, a fast or memory-efficient assembler allows *de novo* assembly to be performed in a limited computing environment (Table 7).

Table 7. Performance summary of short read assemblers.

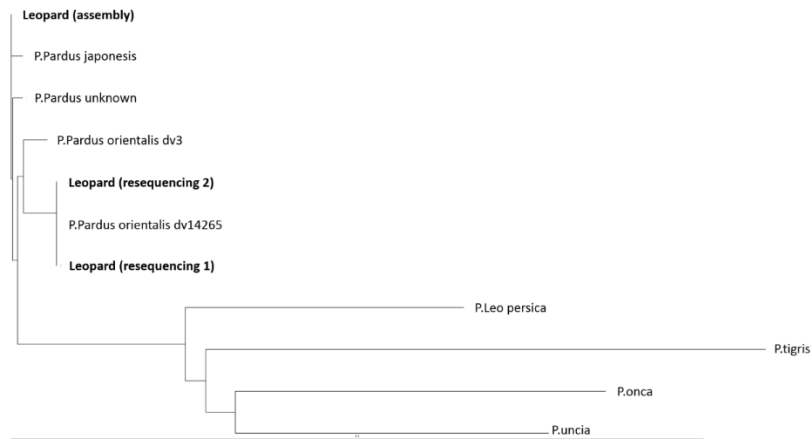
Assembler	Speed	Memory requirement	N50 length
Celera assembler ⁶⁸	slow	high	long
ALLPATHS-LG ⁷²	slow	high	long
ABYSS ⁷³	medium	low	medium
Velvet ⁷⁴	medium	medium	short
SPAdes ⁶⁵	medium	low	medium
SOAPdenovo2 ¹³	fast	medium	medium
SparseAssembler ⁷⁵	medium	low	medium
SGA ⁷⁶	fast	medium	short
MaSuRCA ⁷⁷	slow	high	long

Lastly, we should consider the performance of the assemblers through previous benchmark results. Assemblathon, a competition that is a periodic and collaborative effort to improve and test the numerous assemblers, provides well-organized benchmark result in terms of the performance of the assemblers⁷⁸. In this competition, researchers assemble a given species using several assemblers and benchmark the results. The results of the assemblies and evaluations described in Assemblathon and Genome Assembly Gold-standard Evaluations (GAGE) suggest that one assembler may perform well in one species but not in another species. Therefore, they recommend that use two or more assemblers⁷⁹.

3.2.1 Leopard genome assembly

The leopard genome was assembled from a muscle sample of a female leopard from the Daejeon O-World of Korea (Fig. 3). The extracted DNA was sequenced to 310× average sequencing coverage by Illumina HiSeq sequencer (Table 3). Sequenced reads were assembled using SOAPdenovo2 program into 265,373 contigs (N50 length of 21.0 Kb) and 50,400 scaffolds (N50 length of 21.7 Mb), totaling 2.58 Gb in length (Table 8).

(a) *NADH5* gene



(b) *CYTb* gene

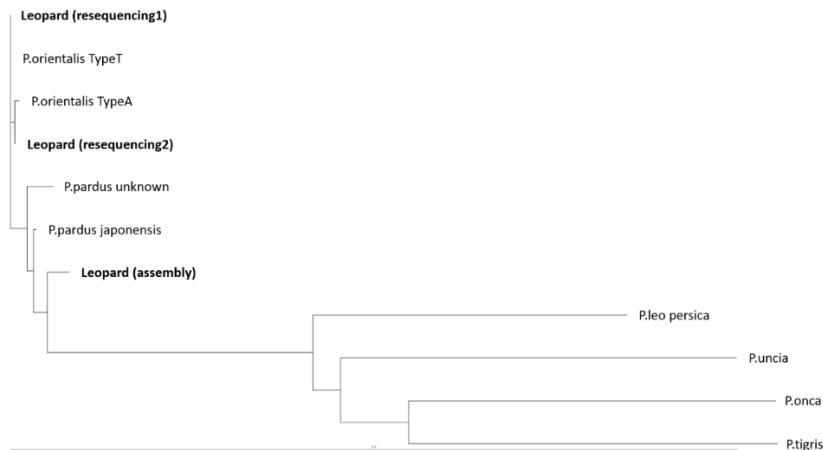


Figure 3. Species and sub-species identification for three leopard samples. (a) *NADH5* and (b) *CYTb* sequences for the three leopards were generated by mapping their reads to the previously reported mitochondrial sequences of *Panthera pardus* (Accession: EF551002.1).

Table 8. The leopard genome assembly statistics

	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	5,500	122,036	3,467,308	135
N80	9,132	87,540	8,979,855	93
N70	12,732	64,613	12,770,773	68
N60	16,634	47,584	18,513,618	51
N50	20,993	34,310	21,701,857	39
Longest	240,914	-----	84,051,066	-----
Total Size	2,478,888,723	-----	2,578,022,254	-----
Total Number (>100bp)	-----	265,235	-----	50,400
Total Number (>2Kb)	-----	174,791	-----	2,670

Additionally, 0.8 coverage of Illumina Truseq long reads (2.0 Gb of total bases) were obtained from two wild Amur leopard individuals (Tables 9 and 10) and were used to fill and correct erroneous gap sequences. The quality evaluation of the leopard genome was performed by comparing the basic statistical values of scaffolds with previously published Felidae genomes and self-aligning the short read to confirm the mapping rate. The leopard genome has been assembled with quality comparable to the previously published Felidae genomes (Table 11). In addition, it was confirmed that about 99% or more short read was aligned well with the leopard genome (Table 12).

Table 9. Sample information of wild Amur leopards and Amur leopard cat used in this study

Species	Sequence	ID	Gender	Data collection	Origin
<i>Panthera pardus orientalis</i>	HiSeq2500	PPO1	M	29 Oct 2006	Nezhenka (Sanduga) river basin , Nadezhdensky Region, Primorsky Krai
<i>Panthera pardus orientalis</i>	TSLR	PPO2	M	02 Nov 2006	Nezhenka (Sanduga) river basin , Nadezhdensky Region, Primorsky Krai
<i>Panthera pardus orientalis</i>	TSLR	PPO4	F	15 Oct 2007	Malaya Ananievka (Elduga) river basin, Nadezhdensky Region, Primorsky Krai
<i>Panthera pardus orientalis</i>	HiSeq2500	PPO5	unknown	18 Oct 2008	Bolshaya Ananievka (Elduga) river basin , Nadezhdensky Region, Primorsky Krai
<i>Prionailurus bengalensis euptilurus</i>	HiSeq2500	-	M	N/A	Republic of Korea

Table 10. Illumina TruSeq Synthetic Long Reads from two wild Amur leopard individuals

# Sequences	393,866
Total bases (bp)	1,999,851,886
Average length (bp)	5,077
Standard deviation (bp)	3,311
The longest length (bp)	21,607
The shortest length (bp)	1,000
N50 (bp)	8,293
GC contents	40.18%
N bases	0.00%

Table 11. Assembly results in five Felidae genomes

	Leopard	Tiger	Cat	Cheetah	Lion
Assembly level	Scaffold	Scaffold	Chromosome	Scaffold	Scaffold
# sequences	50,400	1,479	19 chromosomes + 267,606 unplaced scaffolds	40,077	87,873
Total bases (bp)	2,578,022,254	2,391,082,183	2,641,342,258	2,375,874,546	2,442,522,584
The longest length (bp)	84,051,066	41,607,841	240,380,223	13,046,067	27,160,947
The shortest length (bp)	197	200	152	100	100
Scaffold N50 (bp)	21,701,857	8,860,407	142,431,058	3,121,442	4,005,654
Contig N50 (bp)	20,993	30,032	43,424	28,223	20,046
GC contents	41.71%	41.40%	41.92%	41.30%	41.27%
N bases	3.85%	2.44%	1.58%	1.77%	3.32%

Table 12. Assembly quality assessment using self-alignments

Library		Number of filtered reads	Number of mapped reads	Percentage of mapped reads
170bp	L1	649,639,158	648,380,590	99.81%
	L2	645,441,596	644,116,197	99.79%
400bp	L1	927,631,254	925,675,327	99.79%
500bp	L1	355,755,802	355,021,715	99.79%
700bp	L1	494,678,080	493,757,237	99.81%
	L2	466,939,662	466,051,141	99.81%
2kb	L1	141,024,484	140,465,545	99.60%
	L2	157,681,268	157,027,375	99.59%
	L3	165,113,480	164,509,242	99.63%
5kb	L1	92,125,928	91,543,490	99.37%
	L2	110,644,774	110,020,435	99.44%
	L3	111,490,528	110,871,283	99.44%
10kb	L1	88,451,252	87,990,350	99.48%
	L2	71,257,114	70,805,035	99.37%
	L3	76,850,626	76,376,319	99.38%
15kb	L1	50,274,968	49,404,144	98.27%
	L2	46,902,002	46,350,755	98.82%
	L3	18,748,228	18,630,160	99.37%
	L4	12,188,106	12,110,503	99.36%
20kb	L1	47,273,942	46,217,382	97.77%
	L2	48,418,062	47,879,499	98.89%

3.2.2 Jellyfish genome assembly

The jellyfish genome was different from the leopard. The size of the genome is relatively small, but its complexity is high, I expected that it is difficult to assemble with short reads, such as Illumina sequence. Therefore, my colleagues and I decided to sequence and assemble jellyfish genome using the following hybrid sequencing data: PacBio single molecule real-time sequencing (SMRT) reads, Illumina Truseq long reads, and Illumina short insert-size and mate-pair reads. First, the extracted genomic DNA was sequenced to a 179 \times average sequencing coverage using a PacBio SMRT long reads (30 SMRT cells), as a major sequencing data source for a contig assembly. I obtained 11.4 Kb of median (N50) length of quality filtered PacBio subreads (Fig. 4 and Table 13). I assembled multiple contig sets using the Falcon assembler¹⁵ with the quality filtered PacBio SMRT subreads from a diverse set of read length cutoffs (5 Kb, 6 Kb, 7 Kb, 8 Kb, 9 Kb, 10 Kb, and 12 Kb; Fig. 5 and Table 14).

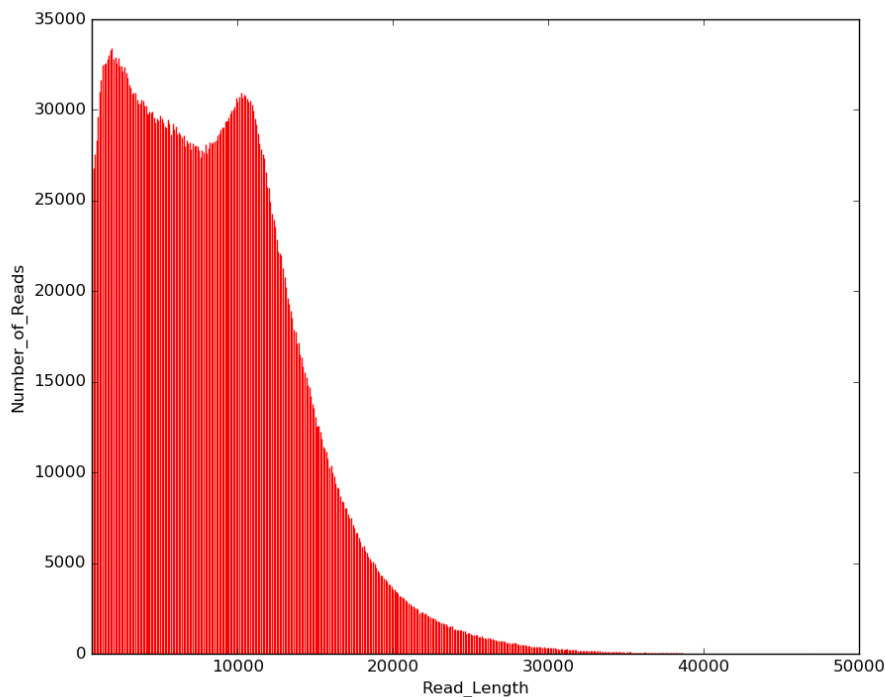


Figure 4. Length distribution of PacBio SMRT reads.

Table 13. PacBio SMRT sequence statistics.

Number of sequences	4,592,385 ea
Total bases	38,170,953,026 bp
Average length	8,311.79 bp
Longest length	50,973 bp
Shortest length	35 bp
N50	11,383 bp
GC contents	38.60%
N bases	0.00%

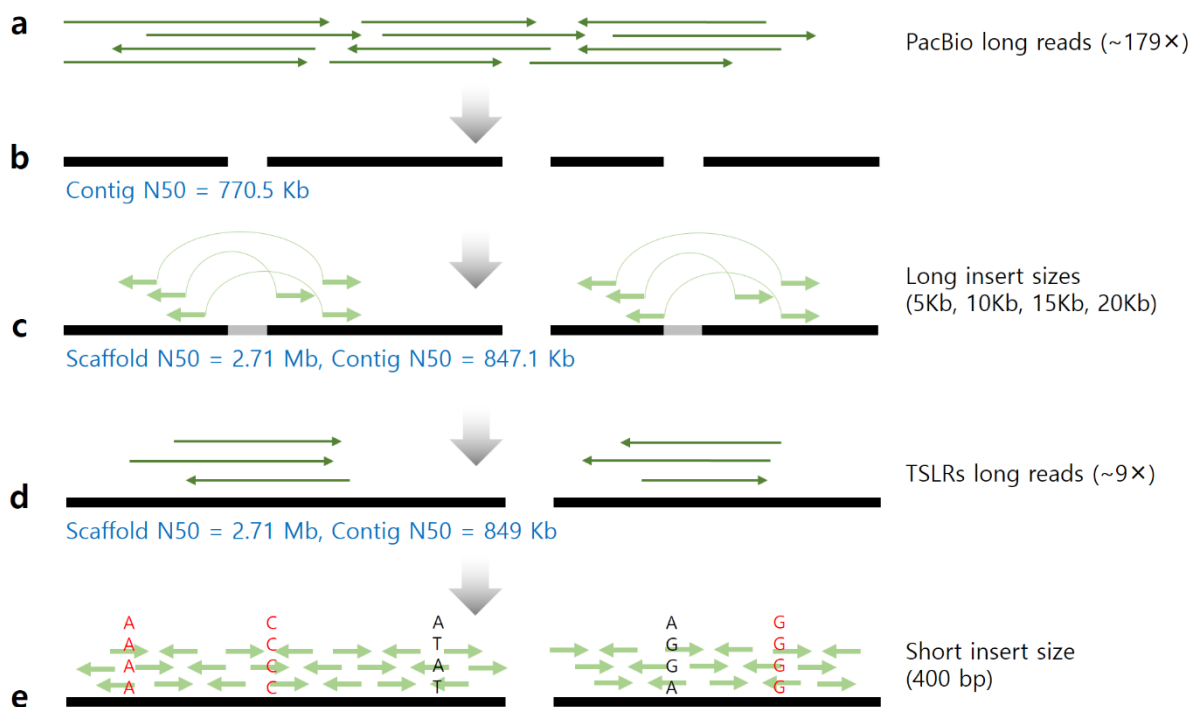


Figure 5. Schematic overview of the *Nemopilema nomurai* genome assembly process. (a) Unassembled PacBio SMRT long reads. **(b)** Contig assembly using PacBio long reads and the Falcon assembler. **(c)** Scaffold assembly using the Illumina mate pair libraries. **(d)** Gap closing using Illumina TruSeq synthetic long reads (TSLR). **(e)** Substitution of common variants using the short insert library. Red denotes common variant that is substituted in the genome.

Table 14. Contig assembly statistics using PacBio SMRT reads.

	PacBio long read length cutoffs						
	5Kb	6Kb	7Kb	8Kb	9Kb	10Kb	12Kb
# of sequences	2,519	2,453	2,078	1,570	1,456	1,140	1,237
Total bases	221,141,034	221,771,871	217,392,668	211,465,427	209,338,243	203,154,934	195,823,825
Longest sequence	3,777,904	3,622,163	3,634,349	5,223,426	4,088,286	4,357,459	2,392,030
Shortest sequence	2	9	2	10	14	10	26
N50	609,640	570,382	669,977	794,113	770,490	952,382	490,833
GC %	38.02%	37.99%	38.07%	38.17%	38.21%	38.20%	38.25%
N bases	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

To extend the contigs into scaffolds, my colleagues and I additionally generated a set of mate-pair libraries (5 Kb, 10 Kb, 15 Kb, and 20 Kb; Table 4). Sequencing and junction adaptor contaminated, low quality (<Q20) and PCR duplicated reads were filtered out and leaving only highly accurate reads for genome assembly. Additionally, short insert size and long insert size reads were trimmed into 90 bp and 50 bp, respectively, to remove low-quality end sequences. I concatenated the contigs to scaffolds using SSPACE¹⁶ and the gaps were filled by aligning the short reads using GapCloser¹³. The scaffold set that was closest to the predicted genome size with the longest N50 length was selected and used for further analyses (Table 15). A total of 255 scaffolds were generated, totaling 213 Mb of sequence length containing only 1.48 % of gaps with an N50 length of 2.71 Mb. Just 92 scaffolds (N90 of 524Kb) successfully covered 90% of the jellyfish genome.

Table 15. Scaffold assembly statistics using PacBio SMRT reads and Illumina mate-pair reads.

	PacBio long read length cutoffs						
	5Kb	6Kb	7Kb	8Kb	9Kb	10Kb	12Kb
Number of sequences	527	464	465	287	255	185	321
Total bases	228,171,285	228,617,968	222,893,641	215,793,878	213,630,333	206,423,756	199,029,964
Longest sequence	7,076,075	5,650,389	6,910,851	6,464,488	8,551,441	11,878,115	3,985,671
Shortest sequence	2	9	2	10	14	10	26
N50	2,266,714	2,149,743	1,759,166	2,209,994	2,711,397	3,064,082	1,204,326
GC %	38.04%	38.00%	38.08%	38.18%	38.23%	38.22%	38.26%
N bases	2.53%	2.45%	1.98%	1.53%	1.48%	1.14%	1.14%

However, genome assemblies constructed using PacBio SMRT reads often contain erroneous sequences (~15%), which are derived from low-quality SMRT reads⁸⁰. Conversely, Illumina TSLRs are generated by local assembly of the high-quality short reads⁸¹. Therefore, I generated 1.92 Gb (~9× coverage) of Illumina TSLRs (Fig. 6 and Table 16) to correct erroneous sequences in the PacBio long-read assembly and to close gap regions. To correct base-pair level errors, I performed three iterations of aligning the Illumina short paired-end sequence to the scaffolds using BWA-MEM¹⁴ and calling variants using SAMtools³⁴. Homozygous variants were substituted using an in-house script. The quality of the assembly was evaluated by aligning the short reads onto the final scaffolds (~99% of mapping rate; Table 17) and by comparing the assembly statistics of other metazoan species. The jellyfish assembly showed the longest assembly continuity among the cnidarian genomes (Table 18).

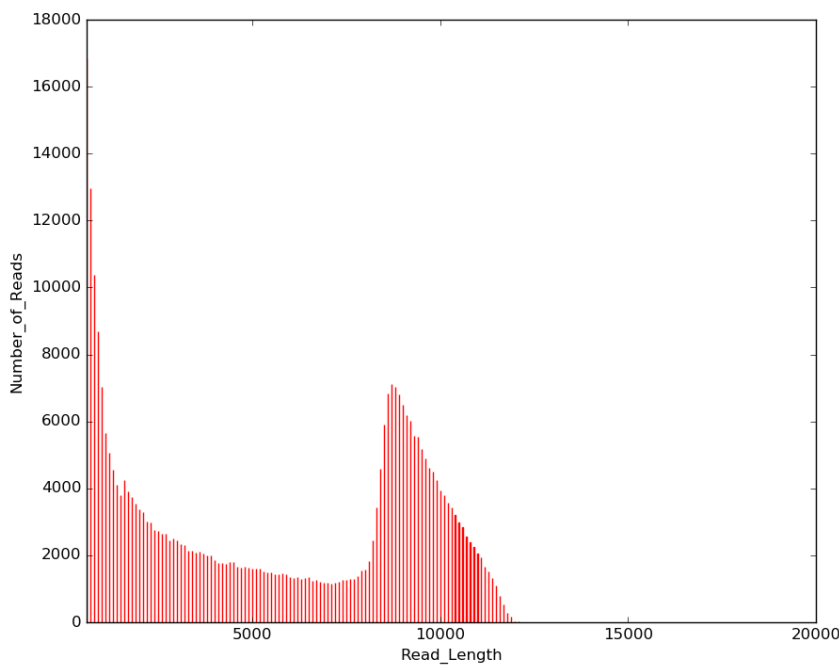


Figure 6. Length distribution of Illumina TruSeq synthetic long reads.

Table 16. Illumina TruSeq Synthetic Long Reads statistics.

Number of sequences	345,790 ea
Total bases	1,922,851,266 bp
Average length	5,560.75 bp
Longest length	20,642 bp
Shortest length	500 bp
N50	8,880 bp
GC contents	38.04%
N bases	0.00%

Table 17. Assembly quality assessment by mapping Illumina reads to *Nemopilema* assembly.

Assembly	Mapping rate				
	400bp	5Kb	10Kb	15Kb	20Kb
<i>Nemopilema</i>	99.74%	99.14%	99.06%	99.02%	98.71%

Table 18. Assembly statistics of nine metazoans and choanoflagellate.

Phylum	Species	NCBI version	# of sequences	Total bases (bp)	Longest length (bp)	Shortest length (bp)	Scaffold N50 (bp)	Contig N50 (bp)	GC ratio	Gap proportion
	<i>Nemopilema nomurai</i>	N/A	255	213,630,333	8,551,441	288	2,711,397	849,297	38.23%	1.48%
	<i>Aurelia aurita</i>	N/A	25454	757,170,055	1,038,510	1001	121,658	14,693	37.48%	12.85%
Cnidaria	<i>Hydra vulgaris</i>	Hydra_RP_1.0	20,916	852,170,992	908,834	2,000	96,317	10,112	27.57%	7.83%
	<i>Clytia hemisphaerica</i>	N/A	7,644	445,210,140	2,888,473	501	366,311	3,860	35.34%	16.63%
	<i>Nematostella vectensis</i>	ASM 20922v1	10,804	356,613,585	3,256,212	626	472,588	19,244	40.64%	16.61%
	<i>Acropora digitifera</i>	Adig_1.1	2,421	447,497,157	2,549,845	2,003	483,559	10,915	39.04%	15.24%
Placozoa	<i>Trichoplax adhaerens</i>	v1.0	1,414	105,631,681	13,260,704	1,000	5,978,658	190,696	32.74%	10.30%
Porifera	<i>Amphimedon queenslandica</i>	v1.0	13,398	166,699,561	1,888,931	633	120,365	11,710	35.83%	13.10%
Ctenophora	<i>Mnemiopsis leidyi</i>	MneLei Aug-11	5,100	155,865,547	1,222,598	987	187,314	11,817	38.86%	3.55%
Holozoa	<i>Monosiga brevicollis</i>	v1.0	219	41,709,928	3,607,471	1,005	1,073,601	48,633	54.81%	7.16%

3.3 Genome annotation

Genome annotation consists of identifying genomic elements, such as, transposable elements and protein-coding genes in the assembly. There are two major gene prediction algorithms: empirical and *ab initio* methods⁸². Also, the empirical method is divided into two types: homology- and evidence-based approaches. The empirical method relies on sequence similarity for detection of homology, while the *ab initio* method uses gene content and signal detection, based on hidden Markov models. In closely-related eukaryotic genomes, empirical algorithms classify DNA regions into coding and non-coding regions based on the assumption that coding regions are more evolutionarily conserved than non-coding regions. If homology cannot be identified with simple sequence alignment, *ab initio* approaches can be used to search the genome for consensus sequences.

The evolutionary position of leopard is very close to that of cat, lion, cheetah and tiger^{2,83-85}. Therefore, I focused on the homology-based and *ab initio* method for leopard gene prediction. A total of 19,043 protein-coding genes were predicted for the leopard genome (Table 19). Additionally, I found that a total of 39.04% of the leopard genome were repetitive elements (Table 20), which is very similar in proportion to the other Felidae species. The GC content and distribution of the leopard genome were also similar to those of the domestic cat and tiger genomes (Fig. 7), indicating little bias in sequencing and assembly.

Table 19. Statistics regarding predicted protein-coding genes in leopard genome

Gene set		Number	Avg. transcript length (bp)	Avg. CDS length (bp)	Avg. no. of exons per gene	Avg. exon length (bp)	Avg. intron length (bp)
<i>De novo</i>	Augustus	22,542	54,517.1	1,455.5	8.9	163.0	6,691.4
	Cat	19,579	47,646.5	1,690.4	10.3	164.8	4,861.2
	Dog	19,890	48,904.2	1,704.6	10.2	166.9	5,000.9
Homolog	Human	20,196	56,969.7	1,732.8	10.3	167.7	5,755.1
	Mouse	22,065	43,752.5	1,655.5	9.6	173.3	4,774.0
	Tiger	18,311	46,718.7	1,671.3	10.5	159.5	4,669.4
Final		19,043	34,265.9	1,618.4	9.4	171.6	3,947.6

Table 20. Statistics regarding transposable elements (TEs) in leopard genome

Type	Ab initio based (bp)	Homology based (bp)	Total (bp)	Percentage of genome (%)
DNA	14,983,643	71,213,110	74,727,059	2.90
LINE	478,274,614	508,794,720	652,640,214	25.32
LTR	42,637,550	126,147,745	131,915,892	5.12
Low_complexity	6,186,939	6,515,731	7,272,828	0.28
SINE	4,687,033	71,002,632	72,009,528	2.79
Satellite	652,202	650,297	1,251,867	0.05
Simple_repeat	44,149,867	44,928,533	48,624,207	1.89
TandemRepeat*			67,553,344	2.62
Unknown	15,308,706	758,634	16,062,533	0.62
Unspecified	389,239		389,239	0.02
Total	605,050,886	829,976,231	1,006,545,511	39.04

* TandemRepeat was separately predicted using TRF program.

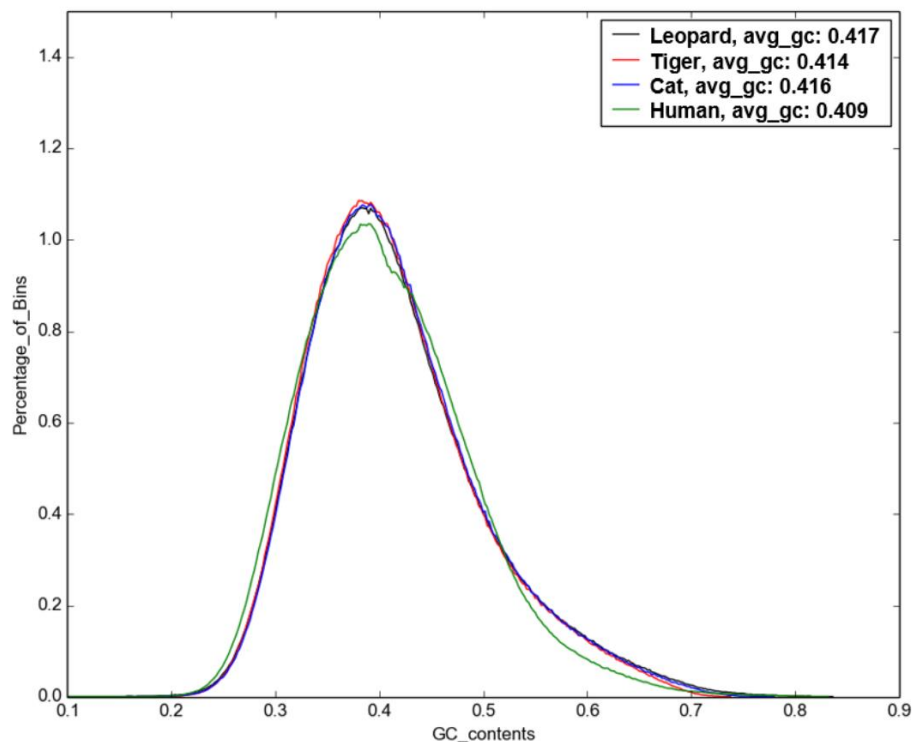


Figure 7. GC content distributions of leopard genome. The x-axis is GC proportion and the y-axis is the proportion of the bin with the specified GC content.

The completeness of genome assembly and annotation was evaluated by the single-copy ortholog mapping approach⁹ (Table 21). The leopard genome showed the highest accuracy and longest continuity among the Panthera species genome assemblies. Two additional wild Amur leopards from the Russian Far East and a wild Amur leopard cat from Korea were also sequenced (Table 22), and were used together with previously reported whole genome sequence data of other Felidae species^{2,83-85} for comparative evolutionary analyses.

Table 21. Assembly and annotation quality assessment of leopard genome using single-copy orthologs mapping approach

	Complete (%)	Duplicated (%)	Fragmented (%)	Missing (%)	Number of single-copy orthologs genes
Leopard	95	0.9	2.5	2.2	3,023
Cat	97	1.3	1.4	0.5	3,023
Cheetah	89	1.3	4.9	5.8	3,023
Lion	87	1.5	5.5	7.2	3,023
Tiger	93	0.6	4.3	2.0	3,023

Table 22. Sequencing statistics regarding two wild Amur leopards and an Amur leopard cat

Sample	# of raw read pairs	# of proper read pairs	% of proper read pairs	Estimated sequencing depth from raw read pairs	Estimated sequencing depth from proper read pairs
Amur leopard-01 (PPO1)	463,914,011	383,291,526	82.62	38.66	31.94
Amur leopard-02 (PPO5)	457,450,100	382,230,035	83.56	38.12	31.85
Amur leopard cat	536,582,305	457,782,689	85.31	44.72	38.15

In the case of jellyfish, the evolutionary distance to hydra, the closest genome to the jellyfish, is about 600 million years²⁸. Therefore, I conduct gene prediction in a different method from the leopard. I applied both empirical (homology- and evidence-based) and ab initio methods. A total of 18,962 protein-coding genes were predicted in jellyfish by combining *de novo* (using tentacle and medusa bell tissue transcriptomes; Table 23) and empirical gene prediction methods (Table 24). The quality assessment of jellyfish assembly and annotation showed the highest recovery rates of single-copy orthologous genes⁹ among all published non-bilaterian metazoan genomes so far (Table 25).

Table 23. Transcriptome sequence statistics of the jellyfish.

Species	Stage	Tissue	Number of raw read pairs	Read length (bp)	Total bases (bp)	Number of clean reads pairs	% of clean reads
<i>Nemopilema nomurai</i>	Medusa	Tentacles	30,909,026	100	6,181,805,200	29,262,691	94.7%
		Bell	33,570,784	100	6,714,156,800	31,656,737	94.3%

Table 24. Statistics of post-filtered protein-coding gene properties in metazoans and holozoan.

Species	# of protein-coding genes	Avg. CDS length (bp)	Avg. exon count	Avg. intron length (bp)	Avg. third codon GC ratio (%)
<i>N. nomurai</i>	18,962	1,441.3	7.5	691.0	0.444
<i>A. aurita</i>	25,174	1,173.9	4.2	1806.1	0.375
<i>H. vulgaris</i>	17,331	1,220.2	5.5	2,612.1	0.246
<i>N. vectensis</i>	24,567	1,003.0	5.3	795.6	0.494
<i>A. digitifera</i>	25,295	1,315.0	6.0	1,118.8	0.420
<i>T. adhaerens</i>	11,491	1,359.6	8.4	283.3	0.310
<i>A. queenslandica</i>	12,811	1,478.9	8.0	263.6	0.376
<i>M. leidyi</i>	15,922	1,385.0	5.5	884.8	0.480
<i>M. brevicollis</i>	9,153	1,801.0	7.5	169.3	0.650
<i>C. elegans</i>	20,256	1,233.5	6.1	307.2	0.405
<i>D. rerio</i>	25,654	1,680.8	9.4	2,796.9	0.547
<i>D. melanogaster</i>	13,864	1,603.7	4.0	973.7	0.639
<i>H. sapiens</i>	19,797	1,735.8	9.9	5,472.2	0.599

Table 25. Gene-set quality assessment of jellyfish using a single-copy ortholog mapping approach.

Species	Complete		Duplicate		Fragment		Missing		Total BUSCO genes
	Count	%	Count	%	Count	%	Count	%	
<i>N. nomurai</i>	409	95.30%	150	35.00%	12	2.80%	8	1.90%	429
<i>A. aurita</i>	323	75.30%	139	32.40%	48	11.20%	58	13.50%	429
<i>H. vulgaris</i>	401	93.50%	129	30.10%	16	3.70%	12	2.80%	429
<i>A. digitifera</i>	342	79.70%	122	28.40%	65	15.20%	22	5.10%	429
<i>N. vectensis</i>	383	89.30%	133	31.00%	29	6.80%	17	4.00%	429
<i>T. adhaerens</i>	397	92.50%	101	23.50%	22	5.10%	10	2.30%	429
<i>A. queenslandica</i>	390	90.90%	124	28.90%	24	5.60%	15	3.50%	429
<i>M. leidy</i>	371	86.50%	88	20.50%	32	7.50%	26	6.10%	429
<i>M. brevicollis</i>	349	81.40%	86	20.00%	34	7.90%	46	10.70%	429
<i>C. elegans</i>	417	97.20%	105	24.50%	4	0.90%	8	1.90%	429
<i>D. rerio</i>	424	98.80%	156	36.40%	4	0.90%	1	0.20%	429
<i>D. melanogaster</i>	425	99.10%	133	31.00%	0	0.00%	4	0.90%	429
<i>H. sapiens</i>	426	99.30%	145	33.80%	2	0.50%	1	0.20%	429

I found that a total of 21.07% of the jellyfish genome was consisted of transposable elements, compared to those of *Hydra vulgaris* (42.87%), *Nematostella vectensis* (33.63%), and *Acropora digitifera* (9.45%) (Table 26). In general, more closely related species are expected to have similar GC contents distribution curves. However, cnidarian species showed very different GC content distributions (Fig. 8). The GC content of *Nemopilema nomurai* is slightly lower than *Acropora digitifera* and *Nematostella vectensis*, but much higher than *Hydra vulgaris*.

Table 26. Repeat annotation of cnidarians.

Repeat type	<i>Nemopilema nomurai</i>	<i>Nematostella vectensis</i>	<i>Acropora digitifera</i>	<i>Hydra vulgaris</i>
DNA	5,440,773	55,668,977	10,897,434	173,628,759
LINE	2,291,406	7,410,687	9,107,195	122,090,336
LTR	1,740,085	8,222,409	7,747,783	7,453,663
Low complexity	269,113	380,420	637,357	6,093,798
Retroposon	-	1,508	-	1,959
SINE	136,032	10,839	49,143	15,574
Satellite	33,340	9,089,245	148,098	130,167
Simple repeat	2,641,456	4,807,011	4,452,246	37,670,762
Tandem repeat	19,010,792	40,720,293	10,842,313	55,194,832
Unknown	27,423,499	3,852,181	973,386	29,189
Unspecified	-	2,621,786	1,206,277	2,967,316
Total TE	45,007,573	119,934,142	42,310,111	365,319,848
Genome size	213,630,333	356,613,585	447,497,157	852,170,992
% of repeat elements	21.07%	33.63%	9.45%	42.87%

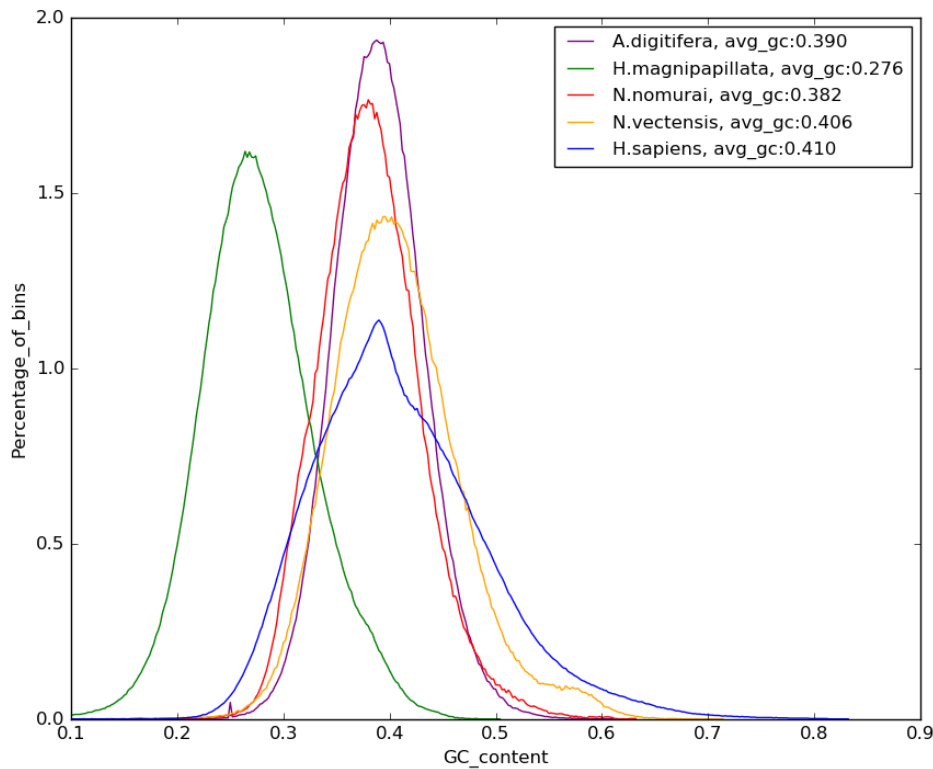


Figure 8. GC content distributions among cnidarian genomes.

3.4 Comparative genomics analysis of leopard and jellyfish

The basis of the evolutionary analysis is comparative genomics, which compares the genome sequences of different species, from bacteria to humans. This process allows researchers to distinguish between different organisms at the molecular level. Also, comparative genomics provides a powerful tool for understanding evolutionary changes between species and helps to identify common or conserved genes among species with genes that give unique characteristics to each species^{86,87}.

According to the evolutionary distance, comparative genomics can be divided into two methods: close species comparative genomics (CSCG) and distant species comparative genomics (DSCG). Close species separated by about 10 million years of evolution (e.g. primates and cats) are especially useful in finding sequence level of differences that can explain differences in phenotype. The CSCG method has been successfully used to compare the genomes of primate, canine, feline, and bovine animals to each other^{2,88-90}. In contrast, very distant species separated by about >1 billion years of evolution cannot compare the sequence differences for the biological features. Therefore, the DSCG and CSCG methods require different approaches. While the CSCG method can analyze positive selection and amino acid changes using sequence differences, the DSCG method can perform the analysis confined to the conserved regions or a part of genes, such as protein domain sequences and the presence/absence of genes.

3.4.1 Comparative genomics analysis of leopard genome

The evolutionary position of leopard is very close to cat, lion, cheetah, and tiger. This enabled comparison with Felidae species as well as with other families, such as Bovidae and Hominidae. I performed tests for deviations in the d_N/d_S ratio (non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site, branch model) and likelihood ratio tests (branch-site model)^{31,32} to detect genes under positive selection for a diet specialized on meat. I found that a total of 586 positively selected genes (PSGs) in the leopard genome. The leopard PSGs were functionally enriched in GTP binding (GO:0005525, 24 genes, $P = 0.00013$), regulation of cell proliferation (GO:0042127, 39 genes, $P = 0.00057$), and macromolecule catabolic process (GO:0009057, 38 genes, $P = 0.00096$; Table 27). Additionally, 228 PSGs were shared in the Felidae family (cat, lion, tiger, cheetah, and leopard); I defined shared PSGs as those that are found in two or more species. The shared PSGs of Felidae were enriched in polysaccharide binding (GO:0030247, 8 genes, $P = 0.00071$), lipid binding (GO:0008289, 12 genes, $P = 0.0041$), and immune response (GO:0006955, 16 genes, $P = 0.0052$; Table 28). Since felid species are hypercarnivores⁹⁰, selection of the lipid binding associated genes may be associated to their obligatory carnivorous diet and regulation of lipid and cholesterol homeostasis^{2,91}.

Table 27. GO enrichment of positively selected genes in leopard

Term	Count	P-value	FDR
GO:0031981~nuclear lumen	71	6.20E-06	0.01
GO:0005654~nucleoplasm	47	5.00E-05	0.07
GO:0031974~membrane-enclosed lumen	81	7.56E-05	0.10
GO:0070013~intracellular organelle lumen	78	9.49E-05	0.13
GO:0043233~organelle lumen	79	1.15E-04	0.16
GO:0005525~GTP binding	24	1.31E-04	0.19
GO:0032561~guanyl ribonucleotide binding	24	1.92E-04	0.28
GO:0019001~guanyl nucleotide binding	24	1.92E-04	0.28
GO:0007264~small GTPase mediated signal transduction	21	3.25E-04	0.56
GO:0042127~regulation of cell proliferation	39	5.72E-04	0.99
GO:0009057~macromolecule catabolic process	38	9.55E-04	1.64
GO:0044265~cellular macromolecule catabolic process	36	9.61E-04	1.65
GO:0006259~DNA metabolic process	27	0.0018	3.10
GO:0000930~gamma-tubulin complex	4	0.0032	4.33
GO:0022613~ribonucleoprotein complex biogenesis	13	0.0043	7.24
GO:0008274~gamma-tubulin ring complex	3	0.0047	6.28
GO:0000931~gamma-tubulin large complex	3	0.0047	6.28
GO:0007049~cell cycle	35	0.0053	8.74

Table 28. GO enrichment of shared positively selected genes in Felidae

Term	Count	P-value	FDR
GO:0044421~extracellular region part	25	6.10E-05	0.08
GO:0009897~external side of plasma membrane	10	7.85E-05	0.10
GO:0005578~proteinaceous extracellular matrix	13	1.45E-04	0.18
GO:0031012~extracellular matrix	13	2.91E-04	0.37
GO:0005539~glycosaminoglycan binding	8	4.04E-04	0.54
GO:0001871~pattern binding	8	7.14E-04	0.96
GO:0030247~polysaccharide binding	8	7.14E-04	0.96
GO:0009986~cell surface	12	0.0011	1.44
GO:0008201~heparin binding	6	0.0032	4.17
GO:0043066~negative regulation of apoptosis	11	0.0038	6.03
GO:0008289~lipid binding	12	0.0041	5.38
GO:0043069~negative regulation of programmed cell death	11	0.0041	6.63
GO:0060548~negative regulation of cell death	11	0.0042	6.76
GO:0007346~regulation of mitotic cell cycle	7	0.0050	7.92
GO:0006955~immune response	16	0.0052	8.22
GO:0005768~endosome	10	0.0062	7.58

If adaptive evolution affects only a few crucial amino acids in a short time interval, none of the measuring selection methods is likely to succeed to define positive selection⁹². Therefore, I investigated target species-specific amino acid changes (AACs) with their effects onto protein function using 15 felines (three leopards, three lions, three tigers, a snow leopard, a cheetah, two leopard cats, and two cats; Table 29) and additional 13 mammalian genomes. It is predicted that 1,509 genes in the felid species had at least one function altering AAC. Unexpectedly but understandably, the Felidae-specific genes with function altering AACs were enriched in DNA repair (GO:0006281, 41 genes, $P = 0.000011$), response to DNA damage stimulus (GO:0006974, 53 genes, $P = 7.39 \times 10^{-7}$), and cellular response to stress (GO:0033554, 63 genes, $P = 0.00016$; Fig. 9; Tables 30 and 31).

Table 29. Variants statistics regarding mapping of Felidae raw reads to the cat reference (Felis_catus_8.0)

Species	All variant sites	Total number of SNV sites	Homozygous SNV sites	Heterozygous SNV sites	Indel sites
Leopard	52,946,286	47,321,889	45,495,382	1,826,507	5,624,397
Amur leopard-01	52,537,072	46,988,478	45,766,378	1,222,100	5,548,594
Amur leopard-02	52,968,234	47,371,008	45,971,258	1,399,750	5,597,226
Lion	50,247,149	45,268,011	41,421,655	3,846,356	4,979,138
Lion-01	52,897,073	47,273,169	45,338,579	1,934,590	5,623,904
White lion	51,618,649	46,195,513	44,564,736	1,630,777	5,423,136
Bengal tiger	51,491,685	45,979,066	43,568,091	2,410,975	5,512,619
Amur tiger	51,057,530	45,861,367	43,157,393	2,703,974	5,196,163
White tiger	48,897,698	43,668,070	41,418,085	2,249,985	5,229,628
Snow leopard	52,483,709	46,887,759	45,770,403	1,117,356	5,595,950
Leopard cat (SRP059496)	38,553,587	34,466,940	28,841,192	5,625,748	4,086,647
Amur leopard cat	42,502,163	37,469,246	32,982,479	4,486,767	5,032,917
Cheetah	36,987,255	32,935,228	31,790,223	1,145,005	4,052,027
Boris cat (SRP039031)	12,295,095	10,512,963	3,609,859	6,903,104	1,782,132

Table 30. GO enrichment of Felidae-specific genes having function altering amino acid changes

Term	Count	<i>P</i> -value	FDR
GO:0000279~M phase	63	1.96E-13	3.53E-10
GO:0022403~cell cycle phase	69	1.07E-11	1.93E-08
GO:0022402~cell cycle process	81	3.25E-10	5.85E-07
GO:0007049~cell cycle	100	8.83E-10	1.59E-06
GO:0006259~DNA metabolic process	72	5.03E-09	9.06E-06
GO:0000087~M phase of mitotic cell cycle	42	6.23E-09	1.12E-05
GO:0000280~nuclear division	41	1.16E-08	2.09E-05
GO:0007067~mitosis	41	1.16E-08	2.09E-05
GO:0048285~organelle fission	41	3.71E-08	6.68E-05
GO:0051301~cell division	47	1.27E-07	2.28E-04
GO:0000793~condensed chromosome	28	1.69E-07	2.44E-04
GO:0006974~response to DNA damage stimulus	53	7.39E-07	0.0013
GO:0043228~non-membrane-bounded organelle	237	1.33E-06	0.0019
GO:0043232~intracellular non-membrane-bounded organelle	237	1.33E-06	0.0019
GO:0044427~chromosomal part	53	2.49E-06	0.0036
GO:0005694~chromosome	60	2.69E-06	0.0039
GO:0000278~mitotic cell cycle	50	6.70E-06	0.012
GO:0051327~M phase of meiotic cell cycle	21	8.97E-06	0.016
GO:0007126~meiosis	21	8.97E-06	0.016
GO:0004518~nuclease activity	27	9.53E-06	0.015
GO:0006281~DNA repair	41	1.10E-05	0.020
GO:0051321~meiotic cell cycle	21	1.23E-05	0.022
GO:0000776~kinetochore	18	1.54E-05	0.022
GO:0005814~centriole	11	2.77E-05	0.040
GO:0000777~condensed chromosome kinetochore	15	2.98E-05	0.043
GO:0005819~spindle	25	6.77E-05	0.098
GO:0000779~condensed chromosome, centromeric region	15	1.35E-04	0.20
GO:0015630~microtubule cytoskeleton	62	1.45E-04	0.21
GO:0033554~cellular response to stress	63	1.61E-04	0.29
GO:0004519~endonuclease activity	18	2.16E-04	0.34
GO:0006310~DNA recombination	19	2.65E-04	0.48
GO:0044450~microtubule organizing center part	13	3.64E-04	0.52
GO:0000723~telomere maintenance	9	4.33E-04	0.78
GO:0032200~telomere organization	9	5.62E-04	1.01
GO:0070193~synaptonemal complex organization	5	6.54E-04	1.17
GO:0007130~synaptonemal complex assembly	5	6.54E-04	1.17
GO:0004896~cytokine receptor activity	12	6.80E-04	1.07
GO:0005739~mitochondrion	103	7.70E-04	1.11
GO:0000775~chromosome, centromeric region	20	8.43E-04	1.21

Table 31. KEGG pathway enrichment of Felidae-specific genes having function altering amino acid changes

Term	Count	<i>P</i> -value	FDR
hsa03450:Non-homologous end-joining	6	6.68E-04	0.81
hsa04060:Cytokine-cytokine receptor interaction	28	0.0040	4.71

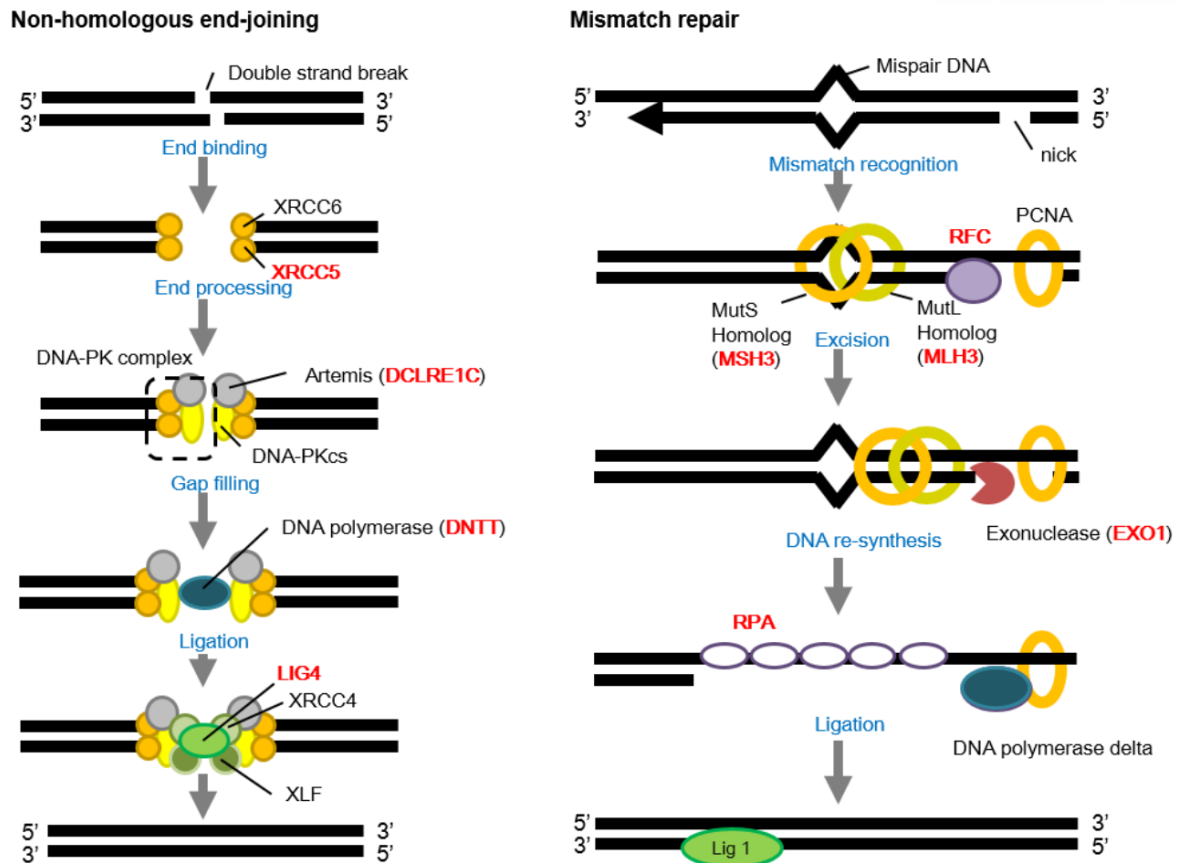


Figure 9. Felidae-specific amino acid changes in DNA repair system. Genes with Felidae-specific function altering amino acid changes in the non-homologous end-joining (KEGG pathway map03450) and mismatch repair (map03430) pathways are shown in red.

Interestingly, three genes (*ACE2*, *MEPIA*, and *PRCP*), which are involved in the protein digestion and absorption pathway, had function altering AACs specific to Felidae species (Figs. 10–12). I interpret this result as a dietary adaptation for high meat consumption that is associated with an increased risk of cancer in humans⁹³, and that the heme-related reactive oxygen species (ROS) in meat cause DNA damage and disrupt normal cell proliferation^{94,95}. I speculate that the functional changes found in DNA damage and repair associated genes help reduce diet related DNA damage in the felid species. This possible felid’s genetic feature can lead to better understanding of human dietary and health research⁹⁶.

Cat	EKANICGMIQGRDDADWVHEDSSQPGQVDHTLVGQCTGAGYFMHFSTR	LGVAEEAALLE	331
Boris cat	EKANICGMIQGRDDADWVHEDSSQPGQVDHTLVGQCTGAGYFMHFSTR	LGVAEEAALLE	331
Leopard cat(2)	EKANICGMIQGRDDADWVHEDSSQPGQVDHTLVGQCTGAGYFMHFSTR	LGVAEEAALLE	331
Cheetah	EKANICGMIQGRDDADWVHEDSSQPGQVDHTLVGQCTGAGYFMHFSTR	LGVAEEAALLE	331
Leopard(3)	EKANICGMIQGRDDADWVHEDSSQPGQVDHTLVGQCTGAGYFMHFSTR	LGVAEEAALLE	331
Lion(3)	EKANICGMIQGRDDADWVHEDSSQPGQVDHTLVGQCTGAGYFMHFSTR	SGVAEEAALLE	331
Tiger(3)	EKANICGMIQGRDDADWVHEDSSQPGQVDHTLVGQCTGAGYFMHFSTR	SGVAEEAALLE	331
Snow leopard	EKANICGMIQGRDDADWVHEDSSQPGQVDHTLVGQCTGAGYFMHFSTR	SGVAEEAALLE	331
Polar bear	EKTNICGMIQGRDDADWVHEDSTQSGQVDHTLVGRCTGAGYFMHFSTR	SSGVAEEAALLE	380
Killer whale	EKANICGMIQGRDRTANV-----QVPAKQKGAGYFMYFSTSLGIAEEAALLE		318
Human	EKANICGMIQGRDRTDWAHQDSAQAGEVDHTLLGQCTGAGYFMQFSTSSGSAEEAALLE		359
Mouse	EKTNVCGMIQGRDDADWAHGDSQPEQVDHTLVGQCKGAGYFMFNTSLGARGEAALE		343
Dog	EKANICGMIQGRDDADWVHEDSTKPGQVDHTLVGRCTGAGYFMHFSTR	SSGMAEEAALLE	331
Pig	EKANICGMIQGRDDADWVHEDSAQPGQVDHTLVGQCTGAGYFMHFDTR	SSGVAEEAALLE	331
Opossum	EKQNICGMIQGRDDEWDIHKRGDSPGQEDHTLVGKCKEAGYFMYFNTSSGMKEEAALLE		349
Panda	EKTNICGMIQGRDDADWVHEDSTQSGQVDHTLVGRCTGAGYFMQFSTSSGMAEEAALLE		307
Cow	EKANICGMIQGRDRTDWDWHENNAQPGQADHTLAGQCTGAGYFMYLNTSFGAAEDAAMLE		331
Horse	EKTNICGMIQGRDRTDWDWVHEDSSQPEQVDHTLVGQCTGAGYFMHLNTSSGSTEEAALLE		355
Rabbit	EKTNICGMIQGRDDADWVRENSDVAGQVDHTLEGQCTGAGYFMHFNTSVGAAEEAALLE		331
Elephant	EKANICGMIQGTIDDADWVHENSVPQVDHTLVGRCTGAGYFMHFSTR	SSGSAEEAALLE	331
Cat	EGTGKGLLEKALPSNLDQEQPSRPKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGFCVN		690
Boriscat	XXXXKXXXXXXXXXNLDQEQPSRXXKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGFCVN		690
Leopardcat(2)	EGTGKGLLEKALPSNLDXGQPSRPKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGICVN		690
Cheetah	EGTGKGLLEKALPSNLDQGQPSRPKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGICVN		690
Leopard(3)	EGTGKGLLEKALPSNLDQGQPSRPKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGICVN		690
Lion(3)	EGTGKGLLEKALPSNLDQGQPSRPKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGICVN		690
Tiger(3)	EGTGKGLLEKALPSNLDQGQPSRPKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGICVN		690
Snow leopard	EGTGKGLLEKALPSNLDQGQPSRPKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGICVN		690
Polar bear	EGSGKVLLEKALPSSLDQGQPGRQKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGICVN		739
Killer whale	EGSGKASLEKDLLGSLGQHRSRQKRSVDNTGPMEDHNWPQYFRDPCDPNQCNEGICVN		677
Human	EGSGKAMLEALPVSLSQGQPSRQKRSVENTGPLEDHNWPQYFRDPCDPNQCNDGICVN		715
Mouse	ESSRKAMLEESLPSSLGQRHPSRQKRSVENTGPMEDHNWPQYFRDPCDPNQCNEGTCVN		701
Dog	AS-EKVSLENALPGSLDQEQPSRQKRSVENTGPLEDHNWPPYFRDPCDPNQCNEGICVN		689
Pig	EASRKASLEKALLESPAQGHSGRQKRSVDNTGPLEDPTWPQYFRDPCDPNQCNEGICVN		690
Opossum	EDSKTMPLLEETQPGSLDA-RPGRQKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGICVN		707
Panda	EGLGKVLLEKALPGSLDQGQPGRQKRSVENTGPLEDHNWPQYFRDPCDPNQCNEGICVN		666
Cow	EDSRKAPLEALPSSLARGQSSRQKRSVDNTGPLXDNWPQYFRDPCDPNQCNEGICVN		690
Horse	EGSGKALLEKALPDSLQGGQSGRQKRSVENTGSLDHNWPQYFGDPNQCNEGICVN		713
Rabbit	AGSEKTLSEAAVAGSLGQGQPSRQKRSVENTGPMEDHKWPQYFRDPCDPNQCNEGICVN		690
Elephant	EATGKALVEALPSSLGQGPSRQKRSVENMSPMEDHNGPQYFRDPCDPNQCNDGTCVN		690

Figure 10. Felidae-specific amino acid change in MEP1A protein. Red rectangles indicate Felidae (2 cats, 2 leopard cats, 1 cheetah, 3 leopards, 3 lions, 3 tigers, and 1 snow leopard)-specific amino acid changes.

Cat	QTIPFVEDNVWSNLKPRISFNFFVTAS	KNVSDVIPRSEVEEAIRMSRSRINDAFRLDDN	720
Boris cat	QTIPFVEDNXWWSNLKPRISFNFFVTAS	KNVSDVIPRSEVEEAIRMSRSRINDAFRLDDN	720
Leopard cat(2)	QTIPFVEDNVWSNLKPRISFNFFVTAS	KNVSDVIPRSEVEEAIRMSRSRINDAFRLDDN	720
Cheetah	QMIPFVEDNVWSNLKPRISFNFFVTAS	KNVSDVIPRSEVKEAIRMSRSRINDAFRLDDN	719
Leopard(3)	QTIPFVEDNVWSNLKPRISFNFFVTAS	KNVSDVIPRREVEEAIRMSRSRINDAFRLDDN	720
Lion(3)	QTIPFVEDNVWSNLKPRISFNFFVTAS	KNVSDVIPRREVEEAIRMSRSRINDAFRLDDN	712
Tiger(3)	QTIPFVEDNVWSNLKPRISFNFFVTAS	KNVSDVIPRREVEEAIRMSRSRINDAFRLDDN	712
Snow leopard	QTIPFVEDNVWSNLKPRISFNFFVTAS	KNVSDVIPRREVEEAIRMSRSRINDAFRLDDN	712
Polar bear	QMIPFVEDNVWVDLKPRI SFNFVTSPGNVSDVIPRADVEGAIKMSRDRINDAFQLDDN		705
Killer whale	KTIPFGEKDVWVSDLKPRISFNFFVTSPKNMSDIIPRTEVEEAIRMSRGRINDAFRLDDN		719
Human	QMILFGEEDVRVANLKPRI SFNFVTAPKNVSDIIPRTEVEKAIRMSRSRINDAFRLDDN		720
Mouse	QTVPFLEEDVRVSDLKPRVSYFFVTSPQNVSDVIPRSEVEDAIRMSRGRINDVFLGNDN		720
Dog	QTIPFVEDNVWVSDLKPRISFNFSVTSPGNVSDIIPRTEVEEAIRMYRSRINDVFLDDN		719
Pig	ETIPFGAVDVWVSDLKPRISFNFFVTSPANMSDIIPRSDVEKAISMSRSRINDAFRLDDN		720
Opossum	QSIPFSNENVKMFDLKPRISFYFFVTSPNGTSFVPREVEEAISMSRDRINDAFRLDDN		719
Panda	QTIPFVEDNVWSNLKPRISFNFFVTSPGNVSDVIPRADVEGAIKMSRDRINDAFRLDDN		720
Cow	ETVLFGEDNVWVSDKKPRISFKFFVTSPNNVSDIIPRTEVENAIRLSRDRINDVFLDDN		719
Horse	QTILFGEEDVRVSDLKPRISFNFFVTSPKNASDIIPRTDVEEAIRMSRSRINDAFRLDDN		720
Rabbit	QTILFGEEDVRVSDLKPRISFNFFVTAPNNVNDIIPRNEVEEAISMSRSRINDIFRLDDN		720
Elephant	QTILFGEEDVWVSDLKPRISFNFFVTVPKNASDIIPKAEVEEAIRMSRGRINDAFRLDDK		715

Figure 11. Felidae-specific amino acid change in ACE2 protein. Red rectangle indicates Felidae (2 cats, 2 leopard cats, 1 cheetah, 3 leopards, 3 lions, 3 tigers, and 1 snow leopard)-specific amino acid changes.

Cat	KDSRYLN	NYLTSEQALADFAVLIKYLKRTIPGAKNQPVIALGGSYGGMLAAWFRMKYPHMV	207
Boris cat	KDSRYLN	NYLTSEQALADFAVLIKYLKRTIPGAKNQPVIALGGSYGGMLAAWFRMKYPHMV	207
Leopard cat(2)	KDSRYLN	NYLTSEQALADFAVLIKNLKRTIPGAKNQPVIAVGGSYGGMLAAWFRMKYPHMV	207
Cheetah	KDSRYLN	NYLTSEQALADFAVLIKYLKRTIPGAKNQPVIALGGSYGGMLAAWFRMKYPHMV	207
Leopard(3)	KDSRYLN	NYLTSEQALADFAVLIKYLKRTIPGAKNQPVIALGGSYGGMLAAWFRMKYPHMV	207
Lion(3)	KDSRYLN	NYLTSEQALADFAVLIKYLKRTIPGAKNQPVIALGGSYGGMLAAWFRMKYPHMV	207
Tiger(3)	KDSRYLN	NYLTSEQALADFAVLIKYLKRTIPGAKNQPVIALGGSYGGMLAAWFRMKYPHMV	196
Snow leopard	KDSRYLN	NYLTSEQALADFAVLIKYLKRTIPGAKNQPVIALGGSYGGMLAAWFRMKYPHMV	196
Polar bear	KDSRHLN	FLTSEQALADFAVLIKHLKRTIPGAKNQPVIAIGGSYGGMLAAWFRMKYPHMV	171
Killer whale	KDSRHLN	FLTTEQALADFAVLIKYLKRTIPGARNQPVIAGGSYGGMLAAWFRMKYPHLV	196
Human	KDSRHLN	FLTSEQALADFAELIKHLKRTIPGAENQPVIAGGSYGGMLAAWFRMKYPHMV	217
Mouse	KDSQH	LNFLTSEQALADFAELIRHLEKTIPGAQQQPVIAGGSYGGMLAAWFRMKYPHIV	194
Dog	KDSRHLN	NYLTSEQALADFAVLIKHLKRTIPGAKNQPVIAIGGSYGGMLAAWFRMKYPHMV	196
Pig	KDSRHLN	FLTSEQALADFAELIRHLKRTIPGTENQPVIAGGSYGGMLAAWFRMKYPHMV	192
Opossum	TDSKHLN	NYLTSEQALADFAELIESLKKNIIPGARNSPVIAIGGSYGGMLAAWFRMKYPHIV	206
Panda	KDSRHLN	FLTSEQALADFAVLIKHLKRTIPGAKNQPVIAVGGSYGGMLAAWFRMKYPHMV	198
Cow	SDSRHLN	FLTTEQALADFAKLI RYLKRTIPGARNQHVIAGGSYGGMLAAWFRMKYPHLV	198
Horse	KDSTHLN	FLTSEQALADFAVLIKHLKRTVPGAKNQPVIALGGSYGGMLAAWFRMKYPHMV	196
Rabbit	KDSRHLN	FLTSEQALADFAELIKHLKRTIPGAENQPVIAGGSYGGMLAAWFRMKYPHVV	132
Elephant	KDSRHLN	FLTSEQALADFAVLIKHLKQTIPGAENQPVIAGGSYGGMLAAWFRMKYPHLV	198

Figure 12. Felidae-specific amino acid change in PRCP protein. Red rectangle indicates Felidae (2 cats, 2 leopard cats, 1 cheetah, 3 leopards, 3 lions, 3 tigers, and 1 snow leopard)-specific amino acid changes.

Conservation of DNA sequences across species reflects functional constraints, and therefore, characterizing genetic variation patterns is critical for understanding the dynamics of genomic change and relevant adaptation of each and a group of species^{97,98}. Homozygous genomic regions are good candidates for evolutionary selection that were need to adapt to environment. I scanned genomic regions that have low level of heterozygous variants, which are strongly conserved among species within three family: Felidae (cat, tiger, cheetah, lion, leopard, snow leopard, and leopard cat, divergence time: ~15.9 million years ago [MYA], carnivores), Hominidae (human, chimpanzee, gorilla, bonobo, and orangutan, ~15.8 MYA, omnivores), and Bovidae (cow, sheep, goat, water buffalo, and yak, ~26 MYA, herbivores)^{28,99,100}. These highly conserved regions (HCRs) represent the reduction in genetic variation (homozygous regions shared among species belonging to the same family; Fig. 13 and Tables 32 and 33).

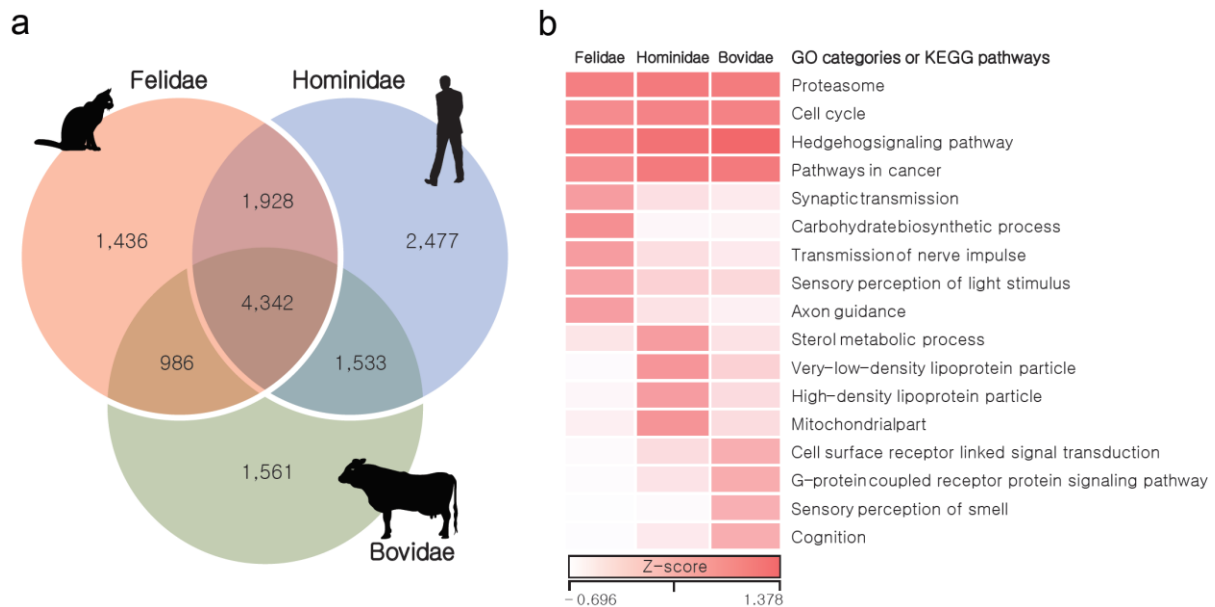


Figure 13. Highly conserved regions in Felidae, Hominidae, and Bovidae. Highly conserved regions in the same family species were identified by calculating the ratios between numbers of conserved and non-conserved positions. **(a)** Venn diagrams of genes in the highly conserved regions. **(b)** Heatmap of enriched gene ontology (GO) categories or KEGG pathways in the highly conserved regions. Z-scores for the average fractions of homozygous positions are shown as a white-to-red color scale.

Table 32. Variants statistics regarding mapping of Felidae raw reads to the cat reference (Felis_catus_8.0)

Species	All variant sites	Total number of SNV sites	Homozygous SNV sites	Heterozygous SNV sites	Indel sites
Leopard	52,946,286	47,321,889	45,495,382	1,826,507	5,624,397
Amur leopard-01	52,537,072	46,988,478	45,766,378	1,222,100	5,548,594
Amur leopard-02	52,968,234	47,371,008	45,971,258	1,399,750	5,597,226
Lion	50,247,149	45,268,011	41,421,655	3,846,356	4,979,138
Lion-01	52,897,073	47,273,169	45,338,579	1,934,590	5,623,904
White lion	51,618,649	46,195,513	44,564,736	1,630,777	5,423,136
Bengal tiger	51,491,685	45,979,066	43,568,091	2,410,975	5,512,619
Amur tiger	51,057,530	45,861,367	43,157,393	2,703,974	5,196,163
White tiger	48,897,698	43,668,070	41,418,085	2,249,985	5,229,628
Snow leopard	52,483,709	46,887,759	45,770,403	1,117,356	5,595,950
Leopard cat (SRP059496)	38,553,587	34,466,940	28,841,192	5,625,748	4,086,647
Amur leopard cat	42,502,163	37,469,246	32,982,479	4,486,767	5,032,917
Cheetah	36,987,255	32,935,228	31,790,223	1,145,005	4,052,027
Boris cat (SRP039031)	12,295,095	10,512,963	3,609,859	6,903,104	1,782,132

Table 33. Variants statistics regarding mapping of Hominidae and Bovidae raw reads to the human and cow references

Family	Species	All variant sites	Total number of SNV sites	Homozygous SNV sites	Heterozygous SNV sites	Indel sites
Hominidae	Bonobo	33,290,642	30,447,841	27,915,325	2,532,516	2,842,801
	Chimpanzee	37,897,572	34,600,658	28,830,656	5,770,002	3,296,914
	Gorilla	45,198,660	41,452,878	36,172,009	5,280,869	3,745,782
	Orangutan	84,426,470	78,815,738	71,088,342	7,727,396	5,610,732
Bovidae	Goat	111,574,672	105,750,483	99,847,134	5,903,349	5,824,189
	Sheep	113,960,484	108,178,988	99,478,910	8,700,078	5,781,496
	Water Buffalo	60,916,988	56,964,575	49,345,127	7,619,448	3,952,413
	Yak	21,285,532	19,538,552	15,873,089	3,665,463	1,746,980

A total of 1.13 Gb of Felidae, 0.88 Gb of Bovidae, and 0.93 Gb of Hominidae HCRs were detected with significantly reduced genetic variation (adjusted $P < 0.0001$, Fisher's exact test corrected using the Benjamini-Hochberg method; Table 34) compared against other genomic regions. Among these regions, a total of 4,342 genes in the HCRs were shared in all three families, and these genes were enriched in many key biological functions (cell cycle, proteasome, pathways in cancer, and Hedgehog signaling pathway; Fig. 13; Tables 35 and 36) as expected. Then, I investigated family-specific genes (1,436 in Felidae, 1,561 in Bovidae and, 2,477 in Hominidae) in the HCRs. The Felidae-specific genes were significantly enriched in synaptic transmission (GO:0007268, 33 genes, $P = 0.0044$), sensory perception of light stimulus (GO:0050953, 27 genes, $P = 0.0022$), axon guidance pathway (20 genes, $P = 0.0054$; Tables 37 and 37), transmission of nerve impulse (GO:0019226, 37 genes, $P = 0.0054$), hinting to adaptation for the fast reflexes found in cats. Interestingly, the Felidae-specific genes were also functionally enriched for carbohydrate biosynthetic process (GO:0016051, 18 genes, $P = 0.00061$). This may be related to the predatory feeding pattern of felids (a meat-based diet, so low dietary availability of carbohydrates).

Table 34. Statistics regarding highly conserved regions in Felidae, Hominidae, and Bovidae genomes

Family	Reference genome size (excluding unplaced fragments)	The number of windows (>80% of sufficiently covered)		Highly conserved windows (Adjusted P -value < 0.0001)		
		Window count	Non-overlapped length (bp)	Window count	Non-overlapped length (bp)	Percentage
Felidae	2,419,212,910	236,332	2,404,232,357	112,821	1,128,179,303	46.92 %
Hominidae	3,088,269,832	267,977	2,732,432,232	93,165	931,656,495	34.10 %
Bovidae	2,660,906,405	257,230	2,616,313,800	87,923	879,223,575	33.61 %

Table 35. GO enrichment of shared genes in the highly conserved regions of Felidae, Hominidae, and Bovidae. Only GO categories with $P < 1.00E-08$ are shown.

Term	Count	P-value	FDR
GO:0031981~nuclear lumen	476	8.08E-31	1.23E-27
GO:0070013~intracellular organelle lumen	557	2.53E-30	3.87E-27
GO:0031974~membrane-enclosed lumen	576	2.54E-30	3.88E-27
GO:0043233~organelle lumen	564	1.95E-29	2.99E-26
GO:0005654~nucleoplasm	307	5.02E-24	7.67E-21
GO:0030528~transcription regulator activity	470	1.84E-17	3.11E-14
GO:0045449~regulation of transcription	751	5.13E-16	1.05E-12
GO:0044451~nucleoplasm part	195	1.20E-15	1.87E-12
GO:0006350~transcription	618	1.12E-14	2.12E-11
GO:0043232~intracellular non-membrane-bounded organelle	682	1.37E-14	2.09E-11
GO:0043228~non-membrane-bounded organelle	682	1.37E-14	2.09E-11
GO:0005730~nucleolus	229	1.91E-14	2.92E-11
GO:0051603~proteolysis involved in cellular protein catabolic process	215	1.56E-13	2.96E-10
GO:0006357~regulation of transcription from RNA polymerase II promoter	250	2.65E-13	5.02E-10
GO:0044257~cellular protein catabolic process	215	2.77E-13	5.24E-10
GO:0043632~modification-dependent macromolecule catabolic process	206	4.50E-13	8.52E-10
GO:0019941~modification-dependent protein catabolic process	206	4.50E-13	8.52E-10
GO:0030163~protein catabolic process	218	1.32E-12	2.50E-09
GO:0045941~positive regulation of transcription	201	1.89E-12	3.58E-09
GO:0045893~positive regulation of transcription, DNA-dependent	175	3.75E-12	7.10E-09
GO:0016568~chromatin modification	114	4.37E-12	8.27E-09
GO:0051254~positive regulation of RNA metabolic process	175	8.43E-12	1.60E-08
GO:0051276~chromosome organization	176	9.51E-12	1.80E-08
GO:0010628~positive regulation of gene expression	203	1.18E-11	2.23E-08
GO:0010604~positive regulation of macromolecule metabolic process	279	1.23E-11	2.33E-08
GO:0044265~cellular macromolecule catabolic process	242	2.17E-11	4.10E-08
GO:0003677~DNA binding	643	3.42E-11	5.77E-08
GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	213	4.14E-11	7.84E-08
GO:0003700~transcription factor activity	303	4.91E-11	8.29E-08
GO:0010557~positive regulation of macromolecule biosynthetic process	221	4.98E-11	9.43E-08
GO:0007049~cell cycle	254	6.33E-11	1.20E-07
GO:0051173~positive regulation of nitrogen compound metabolic process	217	1.02E-10	1.94E-07
GO:0022402~cell cycle process	192	6.33E-10	1.20E-06
GO:0009891~positive regulation of biosynthetic process	227	9.19E-10	1.74E-06
GO:0031328~positive regulation of cellular biosynthetic process	224	1.07E-09	2.03E-06
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	135	2.47E-09	4.68E-06
GO:0009057~macromolecule catabolic process	248	2.78E-09	5.26E-06
GO:0006325~chromatin organization	136	4.91E-09	9.30E-06
GO:0005829~cytosol	359	5.49E-09	8.39E-06

Table 36. KEGG pathway enrichment of shared genes in the highly conserved regions of Felidae, Hominidae, and Bovidae

Term	Count	<i>P</i> -value	FDR
hsa04110:Cell cycle	45	1.94E-04	0.24
hsa05200:Pathways in cancer	97	2.52E-04	0.31
hsa05211:Renal cell carcinoma	28	6.68E-04	0.83
hsa03050:Proteasome	21	7.75E-04	0.96
hsa04340:Hedgehog signaling pathway	23	0.0016	1.92
hsa04120:Ubiquitin mediated proteolysis	45	0.0018	2.18
hsa03018:RNA degradation	23	0.0020	2.50
hsa04914:Progesterone-mediated oocyte maturation	30	0.0047	5.74
hsa04114:Oocyte meiosis	36	0.0059	7.12
hsa00230:Purine metabolism	47	0.0059	7.13

Table 37. GO enrichment of Felidae-specific genes in the highly conserved regions

Term	Count	P-value	FDR
GO:0006811~ion transport	84	4.56E-06	0.008
GO:0005261~cation channel activity	40	1.10E-05	0.018
GO:0046873~metal ion transmembrane transporter activity	45	1.30E-05	0.021
GO:0016892~endoribonuclease activity, producing 3'-phosphomonoesters	9	1.79E-05	0.029
GO:0016894~endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 3'-phosphomonoesters	10	2.39E-05	0.038
GO:0005216~ion channel activity	49	4.01E-05	0.064
GO:0005509~calcium ion binding	95	4.36E-05	0.069
GO:0004522~pancreatic ribonuclease activity	8	4.56E-05	0.073
GO:0004521~endoribonuclease activity	13	5.51E-05	0.088
GO:0022836~gated channel activity	41	7.70E-05	0.12
GO:0022838~substrate specific channel activity	49	8.64E-05	0.14
GO:0006812~cation transport	61	9.06E-05	0.16
GO:0015267~channel activity	50	1.04E-04	0.16
GO:0022803~passive transmembrane transporter activity	50	1.11E-04	0.18
GO:0034702~ion channel complex	30	1.23E-04	0.18
GO:0030001~metal ion transport	53	1.25E-04	0.22
GO:0044459~plasma membrane part	190	1.78E-04	0.26
GO:0034703~cation channel complex	22	2.03E-04	0.29
GO:0031226~intrinsic to plasma membrane	114	2.40E-04	0.35
GO:0005887~integral to plasma membrane	111	3.46E-04	0.50
GO:0004519~endonuclease activity	18	4.14E-04	0.66
GO:0045177~apical part of cell	26	4.28E-04	0.61
GO:0004540~ribonuclease activity	14	4.56E-04	0.72
GO:0016051~carbohydrate biosynthetic process	18	6.13E-04	1.10
GO:0015672~monovalent inorganic cation transport	37	0.0010	1.87
GO:0005886~plasma membrane	297	0.0012	1.74
GO:0050877~neurological system process	107	0.0013	2.39
GO:0016324~apical plasma membrane	20	0.0015	2.19
GO:0034637~cellular carbohydrate biosynthetic process	13	0.0017	3.02
GO:0007601~visual perception	27	0.0022	3.80
GO:0050953~sensory perception of light stimulus	27	0.0022	3.80
GO:0034706~sodium channel complex	6	0.0024	3.37
GO:0022843~voltage-gated cation channel activity	21	0.0024	3.82
GO:0004518~nuclease activity	22	0.0026	4.02
GO:0031224~intrinsic to membrane	412	0.0026	3.65
GO:0007267~cell-cell signaling	58	0.0033	5.81
GO:0031420~alkali metal ion binding	28	0.0037	5.78
GO:0055085~transmembrane transport	55	0.0043	7.45
GO:0007268~synaptic transmission	33	0.0044	7.71
GO:0019226~transmission of nerve impulse	37	0.0054	9.24
GO:0006816~calcium ion transport	19	0.0057	9.80

Table 38. KEGG pathway enrichment of Felidae-specific genes in the highly conserved regions

Term	Count	P-value	FDR
hsa04360:Axon guidance	20	0.0054	6.42

In contrast, the Bovidae-specific genes were enriched in cognition (GO:0050890, 113 genes, $P = 2.54 \times 10^{-9}$; Tables 39–41) and sensory perception of smell (GO:0007608, 82 genes, $P = 2.44 \times 10^{-16}$) functions. I interpreted these functions as a herbivores' adaptation for defense mechanisms from being poisoned by toxic plants¹⁰¹.

Table 39. GO enrichment of Hominidae-specific genes in the highly conserved regions

Term	Count	P-value	FDR
GO:0043235~receptor complex	26	5.65E-04	0.83
GO:0044429~mitochondrial part	91	6.04E-04	0.89
GO:0034364~high-density lipoprotein particle	10	8.15E-04	1.20
GO:0055085~transmembrane transport	89	9.16E-04	1.68
GO:0005887~integral to plasma membrane	160	0.0023	3.33
GO:0005886~plasma membrane	456	0.0026	3.73
GO:0033700~phospholipid efflux	6	0.0027	4.87
GO:0031090~organelle membrane	148	0.0030	4.39
GO:0005789~endoplasmic reticulum membrane	45	0.0035	5.00
GO:0016125~sterol metabolic process	22	0.0035	6.33
GO:0034361~very-low-density lipoprotein particle	8	0.0037	5.42
GO:0034385~triglyceride-rich lipoprotein particle	8	0.0037	5.42
GO:0044432~endoplasmic reticulum part	55	0.0039	5.58
GO:0007155~cell adhesion	102	0.0039	7.05
GO:0022610~biological adhesion	102	0.0041	7.33
GO:0031226~intrinsic to plasma membrane	161	0.0041	5.96
GO:0001570~vasculogenesis	12	0.0044	7.74
GO:0001819~positive regulation of cytokine production	20	0.0045	8.01
GO:0004713~protein tyrosine kinase activity	31	0.0048	7.57
GO:0008092~cytoskeletal protein binding	76	0.0049	7.65
GO:0005739~mitochondrion	145	0.0052	7.45
GO:0005516~calmodulin binding	27	0.0058	9.01
GO:0005740~mitochondrial envelope	63	0.0066	9.35

Table 40. GO enrichment of Bovidae-specific genes in the highly conserved regions

Term	Count	P-value	FDR
GO:0007608~sensory perception of smell	82	2.44E-16	4.00E-13
GO:0007606~sensory perception of chemical stimulus	87	3.36E-16	6.00E-13
GO:0004984~olfactory receptor activity	81	1.36E-15	2.11E-12
GO:0007166~cell surface receptor linked signal transduction	210	5.90E-13	1.06E-09
GO:0007186~G-protein coupled receptor protein signaling pathway	143	1.78E-12	3.19E-09
GO:0007600~sensory perception	111	1.02E-11	1.83E-08
GO:0050890~cognition	113	2.54E-09	4.57E-06
GO:0050877~neurological system process	137	1.70E-08	3.06E-05
GO:0005886~plasma membrane	306	1.55E-04	0.22
GO:0044427~chromosomal part	44	9.19E-04	1.30
GO:0030141~secretory granule	25	0.0011	1.63
GO:0000785~chromatin	26	0.0023	3.21
GO:0043120~tumor necrosis factor binding	5	0.0025	3.82
GO:0000786~nucleosome	12	0.0032	4.44
GO:0005694~chromosome	48	0.0032	4.53
GO:0004499~flavin-containing monooxygenase activity	4	0.0033	5.07
GO:0031091~platelet alpha granule	11	0.0041	5.66
GO:0005576~extracellular region	166	0.0043	5.89
GO:0019932~second-messenger-mediated signaling	29	0.0045	7.74
GO:0016165~lipoxygenase activity	4	0.0062	9.39

Table 41. KEGG pathway enrichment of Bovidae-specific genes in the highly conserved regions

Term	Count	P-value	FDR
hsa04740:Olfactory transduction	83	1.20E-17	1.47E-14

Carnivores tend to have smaller population sizes than species belonging to lower trophic groups, a characteristic argued to be associated with a higher propensity for extinction^{102,103}. I have investigated genetic diversity (which is affected by population size) in Felidae to compare with different dietary requirement groups, herbivores (Bovidae) and omnivores (Hominidae). The Felidae genetic diversity (0.00094 on average), based on the heterozygous single nucleotide variation (SNV) rates, is much lower than those of Bovidae (0.00244) and Hominidae (0.00175; Fig. 14a and Tables 32 and 33). In terms of genomic similarity, Felidae showed the closest genetic distances (0.00102 on average), whereas larger genetic distances were detected in Bovidae (0.00133 on average) and Hominidae (0.00141 on average); suggesting that the extreme dietary specialization in the felids imposes similar and strong selection pressures on its members^{102,103}. The heterozygous SNV rates of leopards (0.00047–0.00070) are similar to those of cheetah (0.00044), snow leopard (0.00043), and white lion (0.00063) that have extremely low genetic diversity due to isolation or inbreeding^{2,85,104}, and smaller than those of tigers (0.00087–0.00104) and lions (0.00074–0.00148). The leopard cats (0.00173–0.00216) show relatively high genetic diversity compared with the larger big cats, as previously reported¹⁰⁵. Additionally, the demographic histories of felid species (leopards, tiger, lion, cheetah, snow leopard, and leopard cat) were constructed using a pairwise sequentially Markovian coalescent (PSMC) model inference⁴⁰. The leopard cat showed a very different demographic history from the big cats: population size of leopard cats increased between 10 million to 2 million years ago, whereas other big cats showed a consistent population decrease (Fig. 14b).

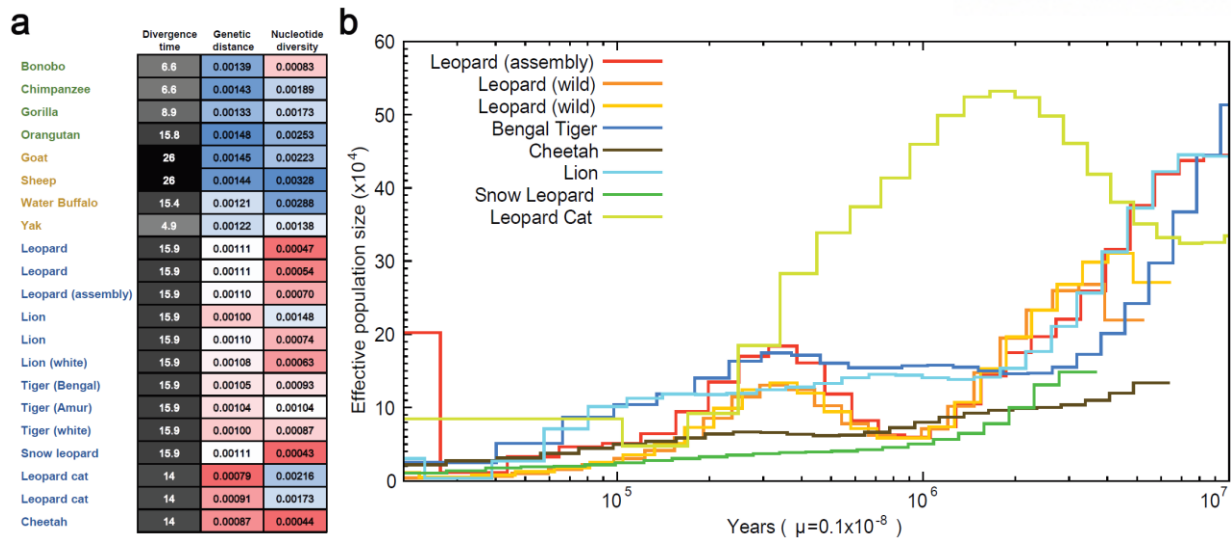


Figure 14. Genetic diversity in Felidae species. (a) Genetic distances and nucleotide diversities. Sequences of Felidae, Hominidae, and Bovidae were mapped to cat, human, and cow references, respectively. The genetic distances were calculated by dividing the number of homozygous SNVs to the reference genome by corresponding species genome size (bp) and divergence time (MYA). Nucleotide diversities were calculated by dividing the number of heterozygous SNVs by the genome size. The divergence times were from TimeTree database. **(b)** Estimated felids population sizes. Generation times of the leopard cat and big cats are 3 and 5 years. μ is mutation rate (per site, per year).

It is predicted that the leopards experienced a strong genetic bottleneck between 2 million to 900 K years ago, whereas other big cats did not. The three leopard genomes showed a similar demographic history. However, over the last 30 K years, the assembled leopard genome showed an explosion in effective population size, whereas the wild leopards did not. The relatively large effective population size likely reflects that admixture occurred very recently between North-Chinese leopard (*P. pardus japonensis*) and Amur leopard, as confirmed by the pedigree information (~30% of North-Chinese leopard admixture) and mitochondrial sequence analyses (Fig. 3), rather than an actual increase in population size. Snow leopard and cheetah and showed low levels of effective population size in the last 3 million years, confirming their low genetic diversity^{2,85}.

3.4.2 Comparative genomics analysis of jellyfish genome

There are limitations in comparison with distant species. Analyzes such as positively selected genes, amino acid change, and highly conserved regions used in the leopard evolutionary analysis are suitable when the nucleotide and amino acid levels can be compared. In the case of jellyfish, the evolutionary distance to moon jellyfish, the closest genome to the *Nemopilema* to date, is about 190 million years. Therefore, comparisons with distantly evolved species commonly use protein domain because nucleotide or amino acid level comparisons are too different in sequence.

I found 20 significantly expanded protein domains in the *Nemopilema* genome. Among them, CUB (PF00431) and Astacin (PF01400) domains are known to be associated with activation of growth factors¹⁰⁶ and regulating development¹⁰⁷, respectively (Fig. 15). Also expanded in *Nemopilema* is the ShK domain-like (PF01549), which is related to Cnidaria toxin¹⁰⁸. These expanded domains were also abundantly found in the previously published *Aurelia aurita* transcriptome study¹⁰⁹.

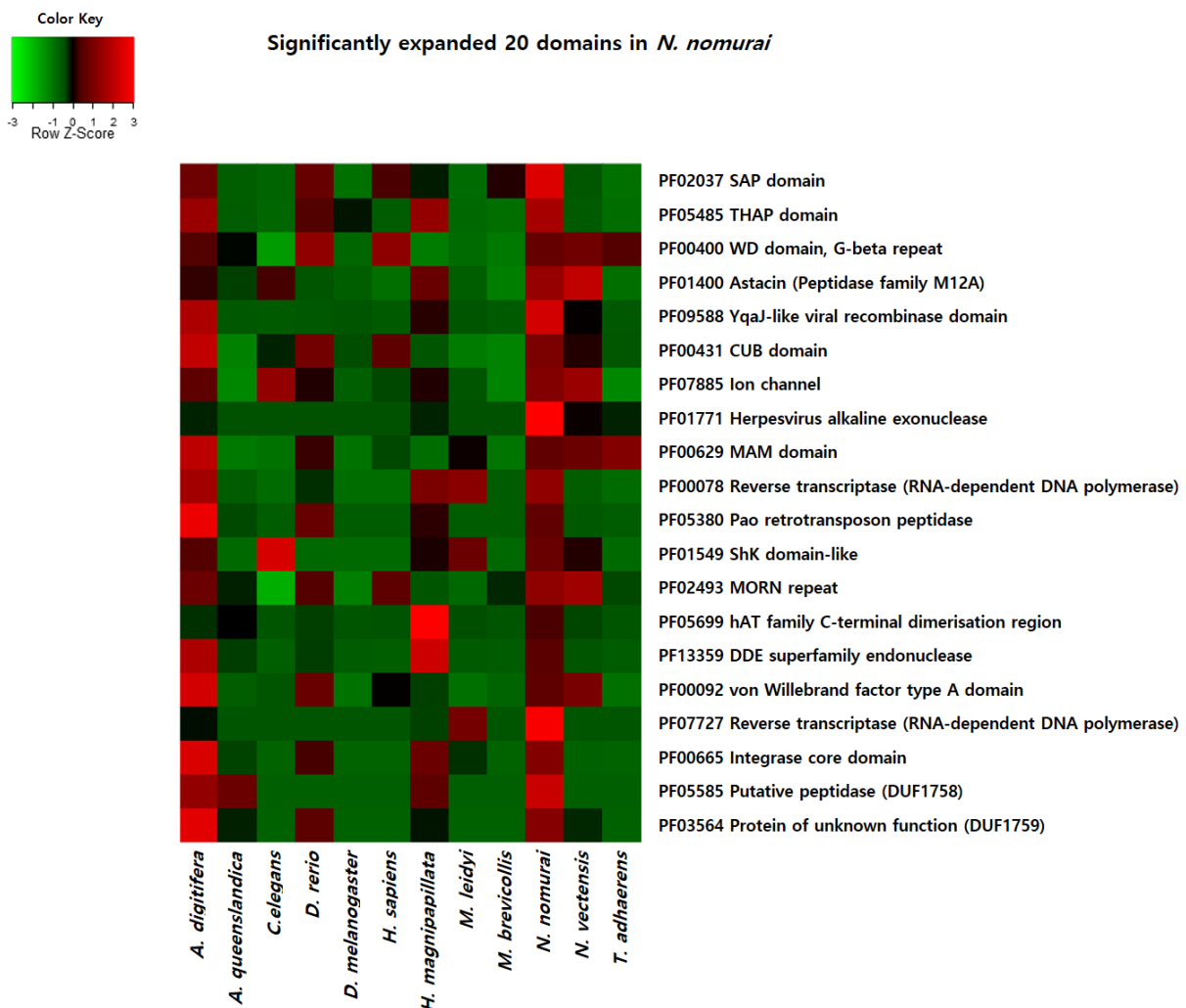


Figure 15. Expanded domains in *Nemopilema nomurai* based on Pfam domain annotation.

A homeobox is a DNA sequence, found within genes that play an important role in the regulation of body plan and morphogenesis in animals, fungi, and plants¹¹⁰. These sequences encode a homeobox domain protein that consists of 60 amino acids helix-turn-helix structure, which is highly conserved among animals. Homeobox genes encode DNA binding protein domains that are involved in the regulation of patterns of anatomical development in animals, and there has been much interest in understanding the early evolution of these genes in the metazoan common ancestor¹¹⁰. There has been much debate surrounding the early evolution of body patterning in the common ancestor of metazoan, particularly concerning the origin and expansion of Hox and *Wnt* gene families^{45,111,112}. In total, 83 homeobox domains were found in *Nemopilema*, while 82, 41, 148, and 120 of homeobox domains were found from *Aurelia*, *Hydra*, *Nematostella* and, *Acropora*, respectively (Table 42).

Table 42. Presence of Hox, Hox-related, and ParaHox homeobox domains in Cnidaria.

Category	Genes	Species				
		<i>Nemopilema</i>	<i>Aurelia</i>	<i>Hydra</i>	<i>Acropora</i>	<i>Nematostella</i>
Hox-related	<i>EVX</i>	O	O	-	O	O
	<i>EMX</i>	O	O	-	O	O
	<i>MOX</i>	O	O	O	O	O
	<i>GBX</i>	-	-	-	O	O
	<i>MNX</i>	-	-	-	O	O
	<i>DLX</i>	O	O	O	O	O
	<i>MSX</i>	O	O	O	O	O
	<i>GSX</i>	O	O	O	O	O
ParaHox	<i>XLOX/CDX</i> (<i>PDX</i>)	O	O	-	-	O
Number of Hox genes		8	7	6	6	7
Total number of homeobox domain		83	82	41	120	148

Interestingly, five of the eight Hox genes in *Nemopilema* are of the posterior type that are associated with aboral axis development¹¹² and clustered with *Nematostella*'s posterior Hox genes, *HOXF*, and *HOXE* (Figs. 16–18). *Aurelia* has six posterior type Hox genes but does not have the *HOXB*, *C*, *D* type (*HOX2* type in humans). Though absent in *Acropora* and *Hydra*, synteny analyses of ParaHox genes in *Nemopilema* show that the *XLOX/CDX* gene is located immediately downstream of *GSX* in the same tandem orientation as those in *Nematostella*, suggesting that *XLOX/CDX* was present in the common ancestor of cnidarian and subsequently lost in some lineages (Fig. 19). Additionally, Hox-related genes, *EVX* and *EMX*, are also present in the scyphozoans (*Aurelia* and *Nemopilema*), although they are lost in *Hydra*.

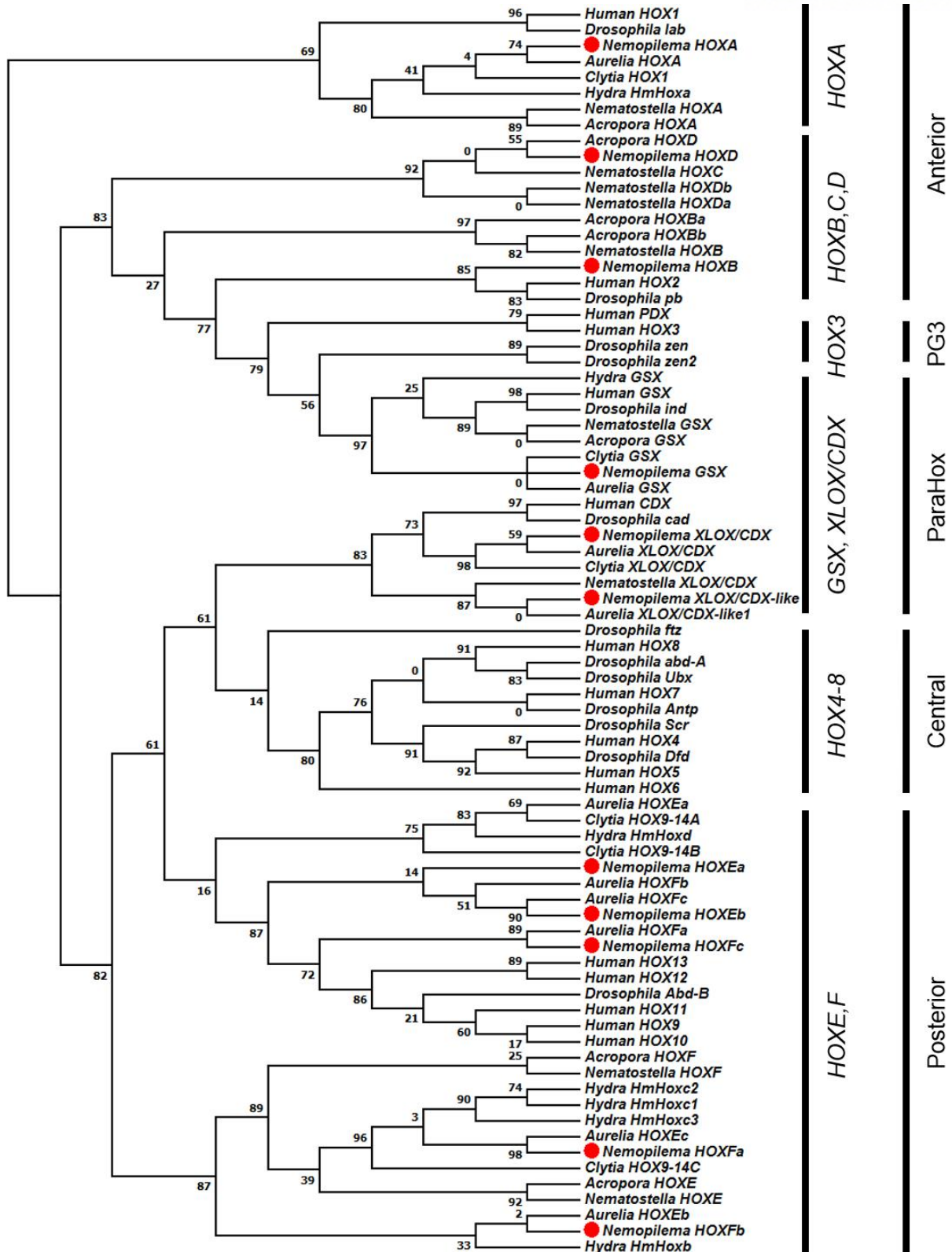


Figure 17. Phylogenetic analysis of Hox and ParaHox homeobox domains with human, fruit fly, and cnidarians. Numbers on nodes denote bootstrap values based on 100 iterations.

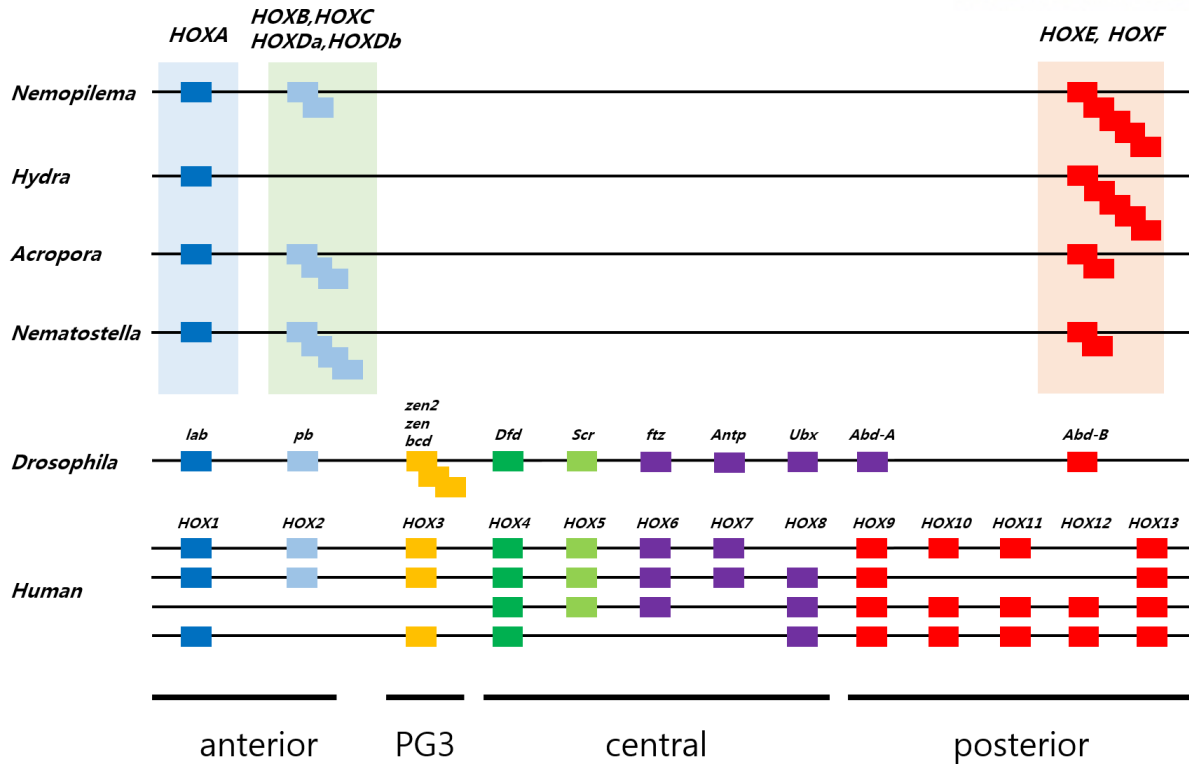


Figure 18. Presence and absence of Hox genes in cnidarians. Blue color indicates anterior Hox genes, yellow color indicates paralogue group 3 (PG3) Hox genes, green and purple colors indicate central Hox genes and red color indicates posterior Hox genes.

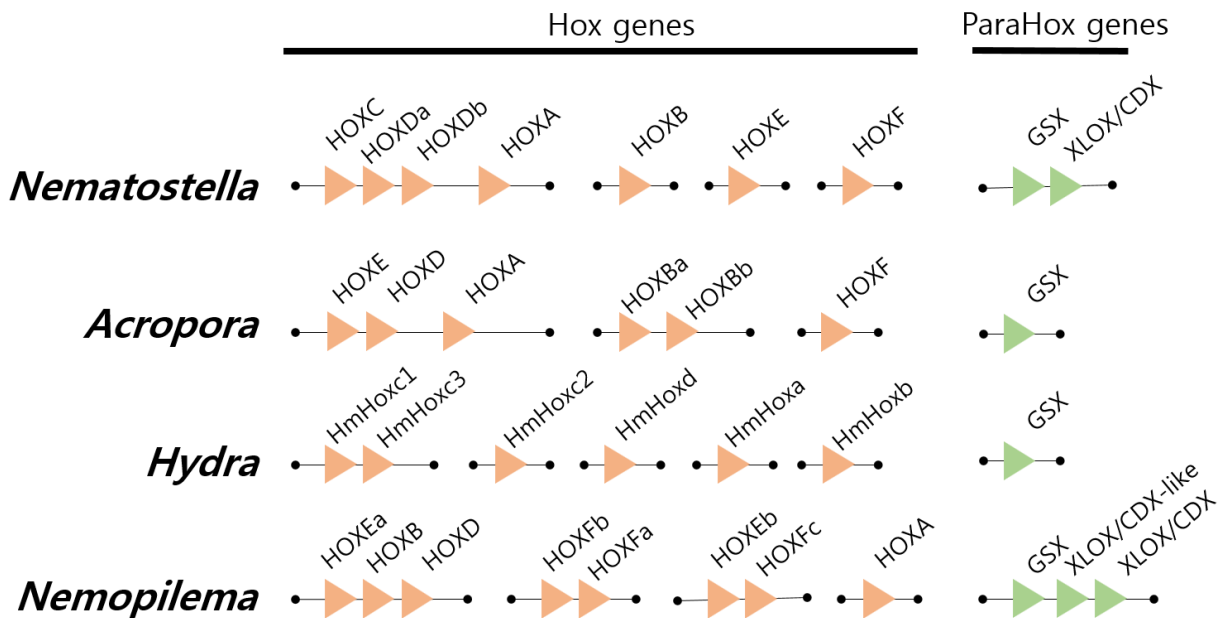


Figure 19. Arrangements of Hox and ParaHox genes in cnidarians. Orange denotes Hox genes, and green denotes ParaHox genes.

Given the large amount of ancestral diversity in the *Wnt* genes, it has been proposed that *Wnt* signaling controlled body plan development in the early metazoans⁴⁸. *Nemopilema* possesses 13 *Wnt* orthologs representing 10 *Wnt* subfamilies (Fig. 20 and Table 43). Notably, *Wnt9* is absent from all cnidarians, likely representing losses in the common ancestor of cnidarian. Interestingly, cnidarians have undergone dynamic lineage-specific *Wnt* subfamily duplications, such as *Wnt8* (*Acropora*, *Nematostella*, and *Aurelia*), *Wnt10* (*Hydra*), and *Wnt11*, and *Wnt16* (*Aurelia* and *Nemopilema*).

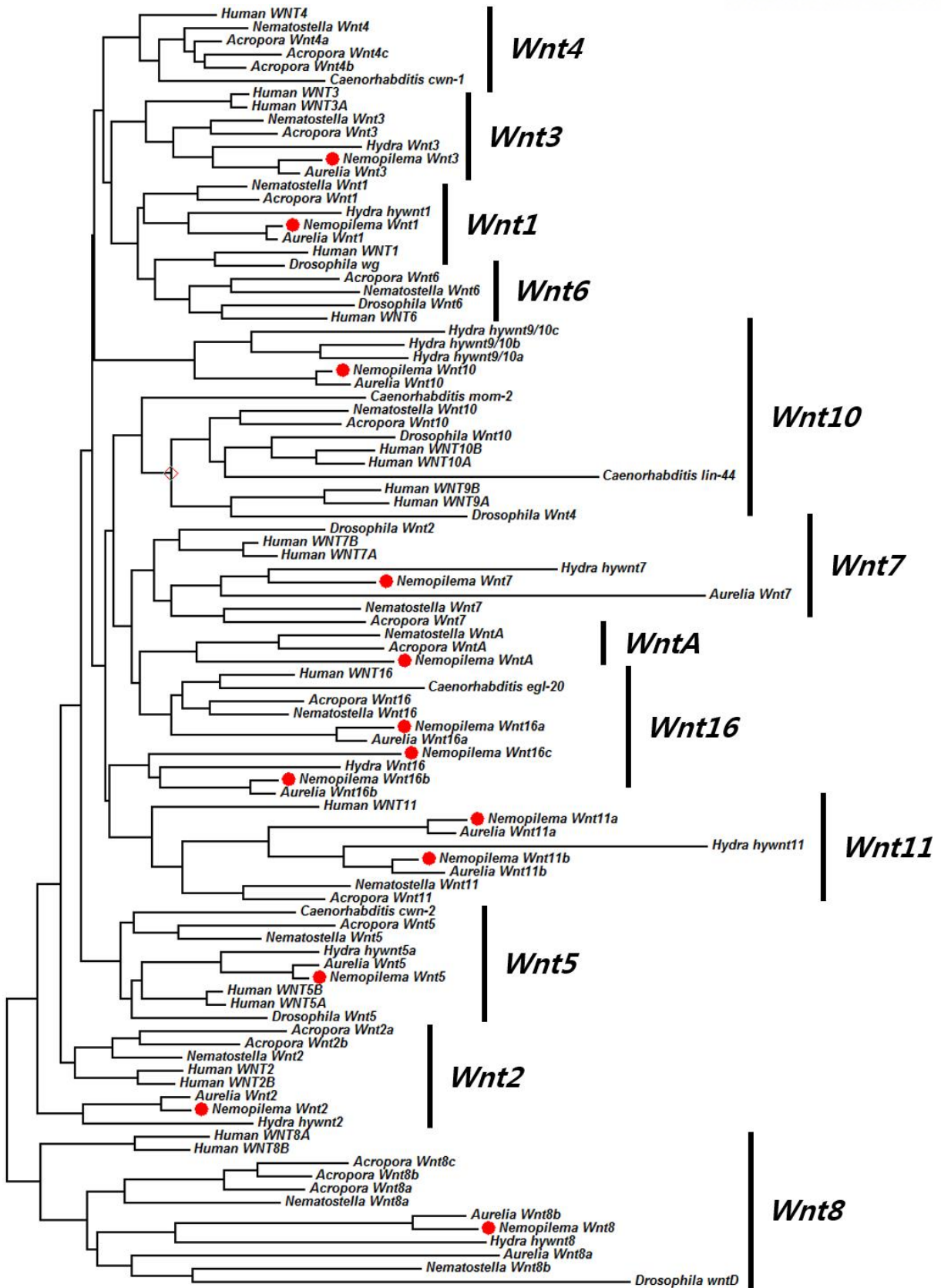


Figure 20. Phylogenetic tree using Maximum likelihood of Wnt proteins.

Table 43. Distribution of *Wnt* genes among cnidarians.

Gene	Cnidaria					Arthropoda	Chordata
	Scyphozoa		Hydrozoa	Anthozoa		Insecta	Mammalia
	<i>Nemopilema</i>	<i>Aurelia</i>	<i>Hydra</i>	<i>Nematostella</i>	<i>Acropora</i>	<i>Drosophila</i>	Human
<i>Wnt1</i>	1	1	1	1	1	1	1
<i>Wnt2</i>	1	1	1	1	2	0	2
<i>Wnt3</i>	1	1	1	1	1	0	2
<i>Wnt4</i>	0	0	0	1	3	0	1
<i>Wnt5</i>	1	1	1	1	1	1	2
<i>Wnt6</i>	0	0	0	1	1	1	1
<i>Wnt7</i>	1	1	1	1	1	1	2
<i>Wnt8</i>	1	2	1	2	3	0	2
<i>Wnt9</i>	0	0	0	0	0	1	2
<i>Wnt10</i>	1	1	3	1	1	1	2
<i>Wnt11</i>	2	2	1	1	1	0	1
<i>Wnt16</i>	3	2	1	1	1	0	1
<i>WntA</i>	1	0	0	1	1	0	0
Other	0	0	0	0	0	1	0
Total	13	12	11	13	17	7	19

It has been proposed that a primordial cluster of *Wnt* genes (*Wnt1–Wnt6–Wnt10*) existed in the last common ancestor of arthropods and deuterostomes¹¹³. Our analyses of cnidarian genomes revealed that *Acropora* also possesses this cluster, while *Aurelia*, *Nemopilema*, and *Hydra* are missing *Wnt6*, suggesting the loss of the *Wnt6* gene in the common ancestor of Medusozoa lineage (Fig. 21). Taken together, the *Nemopilema* has the comparable number of *Wnt* and Hox genes to other cnidarians, but the dynamic repertoire of these gene families suggests that cnidarians have evolved independently to adapt their physiological characteristics and life cycle.

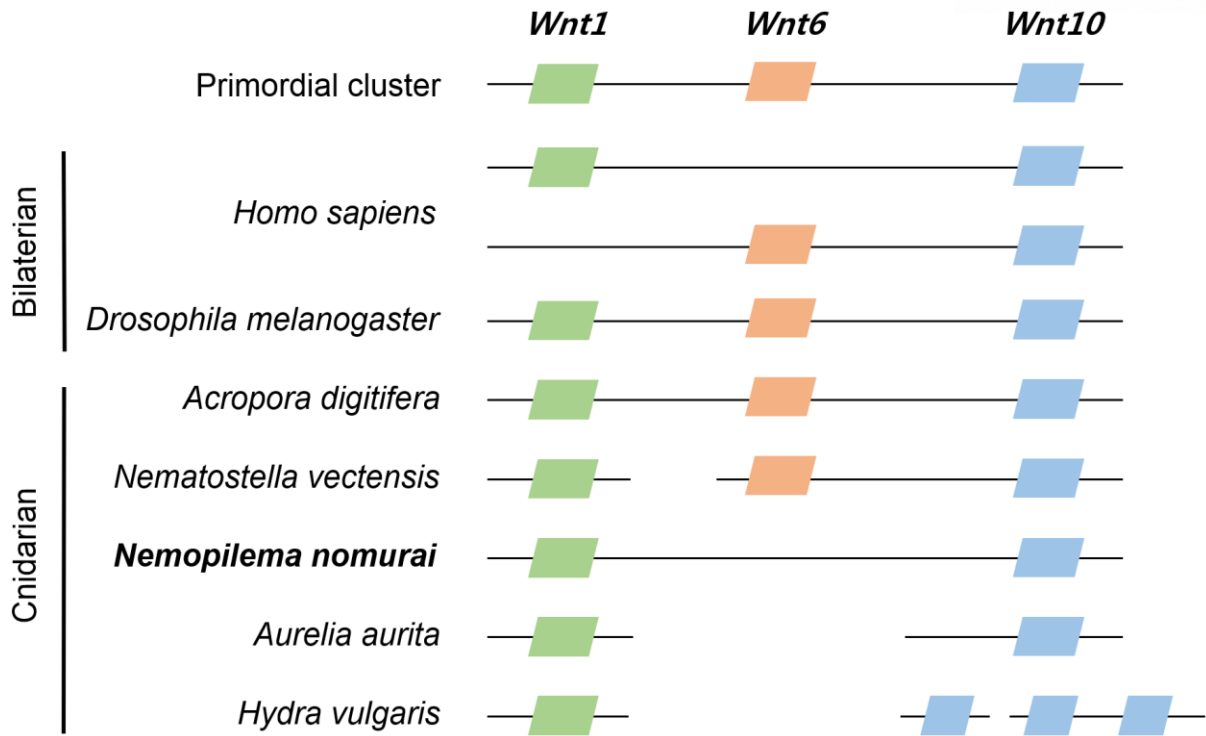


Figure 21. A primordial cluster of three *Wnt* gene (*Wnt1–Wnt6–Wnt10*) pattern of cnidarians.

IV. Conclusions

Limitations of current sequencing techniques and analysis tools poses challenges to the genome assembly. Therefore, it is important to understand the current technologies to establish a proper sequencing and analysis strategy. This study presented guidelines for the sequencing platform, the choice of assembler, the genome characteristics of a species, and comparative analysis strategies based on the presence or absence of closely related species through the leopard and jellyfish genomes.

The first consideration is the selection of the sequencing platform that has the most significant impact on genome assembly quality. Short-read sequencing is a cost-effective method to produce genome assembly, but it has shown poor performance for repetitive or GC-biased regions⁵⁶. Long-read sequencing is useful for resolving those problems, and it provides more continuous assembly than those of the short-read. This process requires high sequencing coverages (>50X) and computational costs to make a high-quality genome⁵⁷. Given current sequencing technology, the ideal method is a hybrid method that sequences a genome with a long-read, scaffolding it with a mate-pair or Hi-C library^{114,115}, and correcting the error of the long-read with a short-read data. This method not only benefits the quality of the genome assembly of the large genome but is more cost effective as well. Choosing the assembler for the *de novo* genome assembly also deserves special consideration. I recommend trying more than one proper assembler considering sequencing platform, genome complexity, computing resources, and performance.

Depending on the evolutionary distance, I suggested two comparative genomics methods: close species comparative genomics (CSCG) and distant species comparative genomics (DSCG). The leopard, evolutionarily proximal to the cat, cheetah, and tiger, has a genome size and GC content graph similar to the cheetah and tiger (Fig. 7 and Table 10). Therefore, previous studies have helped to establish an analysis strategy for the leopard genome. Analyses of positive selection, amino acid changes, and highly conserved regions used in the leopard are basically suitable when the nucleotide and amino acid levels can be compared. In the case of jellyfish, by contrast, the evolutionary distance to moon jellyfish (*Aurelia aurita*), the closest genome to the *Nemopilema* to date, is about 200 million years. Therefore, distant species comparative genomics use the protein domains and absence/presence of conserved genes because nucleotide or amino acid level comparisons are too heterogeneous in their sequences (see Figs. 16 and 17).

In this study, I presented the guidelines for a *de novo* genome assembly by analyzing the leopard and jellyfish genomes. The two genomes showed successful genome assembly with different strategies. Compared to the second- and third-generation sequencing technologies used in this study, the recently released Oxford Nanopore technology can provide high-throughput long reads at an affordable cost⁶¹, and Hi-C technology can be used to complete longer scaffold assembly. Moreover,

Optical maps and Bionano technologies can achieve extended scaffolding with the correction of missassemblies^{1,116}. By combining these technologies, I expect to be able to assemble a high-quality assembly with chromosome level. I think that the development of sequencing technologies will facilitate the discovery of new genomes, as many species have not been unveiled yet.

References

1. Cho Y. S.; Kim H.; Kim H. M.; Jho S.; Jun J.; Lee Y. J.; Chae K. S.; Kim C. G.; Kim S.; Eriksson A.; Edwards J. S.; Lee S.; Kim B. C.; Manica A.; Oh T. K.; Church G. M.; Bhak J. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat Commun* **2016**, *7*, 13637.
2. Cho Y. S.; Hu L.; Hou H.; Lee H.; Xu J.; Kwon S.; Oh S.; Kim H. M.; Jho S.; Kim S.; Shin Y. A.; Kim B. C.; Kim H.; Kim C. U.; Luo S. J.; Johnson W. E.; Koepfli K. P.; Schmidt-Kuntzel A.; Turner J. A.; Marker L.; Harper C.; Miller S. M.; Jacobs W.; Bertola L. D.; Kim T. H.; Lee S.; Zhou Q.; Jung H. J.; Xu X.; Gadhvi P.; Xu P.; Xiong Y.; Luo Y.; Pan S.; Gou C.; Chu X.; Zhang J.; Liu S.; He J.; Chen Y.; Yang L.; Yang Y.; He J.; Liu S.; Wang J.; Kim C. H.; Kwak H.; Kim J. S.; Hwang S.; Ko J.; Kim C. B.; Kim S.; Bayarlkhagva D.; Paek W. K.; Kim S. J.; O'Brien S. J.; Wang J.; Bhak J. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat Commun* **2013**, *4*, 2433.
3. Yim H. S.; Cho Y. S.; Guang X.; Kang S. G.; Jeong J. Y.; Cha S. S.; Oh H. M.; Lee J. H.; Yang E. C.; Kwon K. K.; Kim Y. J.; Kim T. W.; Kim W.; Jeon J. H.; Kim S. J.; Choi D. H.; Jho S.; Kim H. M.; Ko J.; Kim H.; Shin Y. A.; Jung H. J.; Zheng Y.; Wang Z.; Chen Y.; Chen M.; Jiang A.; Li E.; Zhang S.; Hou H.; Kim T. H.; Yu L.; Liu S.; Ahn K.; Cooper J.; Park S. G.; Hong C. P.; Jin W.; Kim H. S.; Park C.; Lee K.; Chun S.; Morin P. A.; O'Brien S. J.; Lee H.; Kimura J.; Moon D. Y.; Manica A.; Edwards J.; Kim B. C.; Kim S.; Wang J.; Bhak J.; Lee H. S.; Lee J. H. Minke whale genome and aquatic adaptation in cetaceans. *Nat Genet* **2014**, *46*, 88-92.
4. Chung O.; Jin S.; Cho Y. S.; Lim J.; Kim H.; Jho S.; Kim H. M.; Jun J.; Lee H.; Chon A.; Ko J.; Edwards J.; Weber J. A.; Han K.; O'Brien S. J.; Manica A.; Bhak J.; Paek W. K. The first whole genome and transcriptome of the cinereous vulture reveals adaptation in the gastric and immune defense systems and possible convergent evolution between the Old and New World vultures. *Genome Biol* **2015**, *16*, 215.
5. Bhak Y.; Jeon Y.; Jeon S.; Chung O.; Jho S.; Jun J.; Kim H. M.; Cho Y.; Yoon C.; Lee S.; Kang J. H.; Lim J. D.; An J.; Cho Y. S.; Ryu D. Y.; Bhak J. Myotis rufoniger genome sequence and analyses: M. rufoniger's genomic feature and the decreasing effective population size of Myotis bats. *PLoS One* **2017**, *12*, e0180418.
6. van Dijk E. L.; Jaszczyszyn Y.; Naquin D.; Thermes C. The Third Revolution in Sequencing Technology. *Trends Genet* **2018**, *34*, 666-681.
7. Miller W.; Drautz D. I.; Janecka J. E.; Lesk A. M.; Ratan A.; Tomsho L. P.; Packard M.; Zhang Y.; McClellan L. R.; Qi J.; Zhao F.; Gilbert M. T.; Dalen L.; Arsuaga J. L.; Ericson P. G.; Huson D. H.; Helgen K. M.; Murphy W. J.; Gotherstrom A.; Schuster S. C. The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). *Genome Res* **2009**, *19*, 213-220.

8. Paszkiewicz K.; Studholme D. J. De novo assembly of short sequence reads. *Brief Bioinform* **2010**, *11*, 457-472.
9. Simao F. A.; Waterhouse R. M.; Ioannidis P.; Kriventseva E. V.; Zdobnov E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210-3212.
10. Chain P. S.; Grafham D. V.; Fulton R. S.; Fitzgerald M. G.; Hostetler J.; Muzny D.; Ali J.; Birren B.; Bruce D. C.; Buhay C.; Cole J. R.; Ding Y.; Dugan S.; Field D.; Garrity G. M.; Gibbs R.; Graves T.; Han C. S.; Harrison S. H.; Highlander S.; Hugenholtz P.; Khouri H. M.; Kodira C. D.; Kolker E.; Kyrpides N. C.; Lang D.; Lapidus A.; Malfatti S. A.; Markowitz V.; Metha T.; Nelson K. E.; Parkhill J.; Pitluck S.; Qin X.; Read T. D.; Schmutz J.; Sozhamannan S.; Sterk P.; Strausberg R. L.; Sutton G.; Thomson N. R.; Tiedje J. M.; Weinstock G.; Wollam A.; Genomic Standards Consortium Human Microbiome Project Jumpstart C.; Detter J. C. Genomics. Genome project standards in a new era of sequencing. *Science* **2009**, *326*, 236-237.
11. Human Microbiome Jumpstart Reference Strains C.; Nelson K. E.; Weinstock G. M.; Highlander S. K.; Worley K. C.; Creasy H. H.; Wortman J. R.; Rusch D. B.; Mitreva M.; Sodergren E.; Chinwalla A. T.; Feldgarden M.; Gevers D.; Haas B. J.; Madupu R.; Ward D. V.; Birren B. W.; Gibbs R. A.; Methe B.; Petrosino J. F.; Strausberg R. L.; Sutton G. G.; White O. R.; Wilson R. K.; Durkin S.; Giglio M. G.; Gujja S.; Howarth C.; Kodira C. D.; Kyrpides N.; Mehta T.; Muzny D. M.; Pearson M.; Pepin K.; Pati A.; Qin X.; Yandava C.; Zeng Q.; Zhang L.; Berlin A. M.; Chen L.; Hepburn T. A.; Johnson J.; McCarrison J.; Miller J.; Minx P.; Nusbaum C.; Russ C.; Sykes S. M.; Tomlinson C. M.; Young S.; Warren W. C.; Badger J.; Crabtree J.; Markowitz V. M.; Orvis J.; Cree A.; Ferriera S.; Fulton L. L.; Fulton R. S.; Gillis M.; Hemphill L. D.; Joshi V.; Kovar C.; Torralba M.; Wetterstrand K. A.; Abouelleil A.; Wollam A. M.; Buhay C. J.; Ding Y.; Dugan S.; FitzGerald M. G.; Holder M.; Hostetler J.; Clifton S. W.; Allen-Verge E.; Earl A. M.; Farmer C. N.; Liolios K.; Surette M. G.; Xu Q.; Pohl C.; Wilczek-Boney K.; Zhu D. A catalog of reference genomes from the human microbiome. *Science* **2010**, *328*, 994-999.
12. Genome K. C. o. S. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered* **2009**, *100*, 659-674.
13. Luo R.; Liu B.; Xie Y.; Li Z.; Huang W.; Yuan J.; He G.; Chen Y.; Pan Q.; Liu Y.; Tang J.; Wu G.; Zhang H.; Shi Y.; Liu Y.; Yu C.; Wang B.; Lu Y.; Han C.; Cheung D. W.; Yiu S. M.; Peng S.; Xiaoqian Z.; Liu G.; Liao X.; Li Y.; Yang H.; Wang J.; Lam T. W.; Wang J. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **2012**, *1*, 18.
14. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997* **2013**.
15. Chin C.-S.; Alexander D. H.; Marks P.; Klammer A. A.; Drake J.; Heiner C.; Clum A.; Copeland

- A.; Huddleston J.; Eichler E. E. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **2013**, *10*, 563-569.
16. Boetzer M.; Henkel C. V.; Jansen H. J.; Butler D.; Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **2011**, *27*, 578-579.
17. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **1999**, *27*, 573-580.
18. Jurka J.; Kapitonov V. V.; Pavlicek A.; Klonowski P.; Kohany O.; Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **2005**, *110*, 462-467.
19. Tarailo-Graovac M.; Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **2009**, 4.10. 11-14.10. 14.
20. Price A. L.; Jones N. C.; Pevzner P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **2005**, *21 Suppl 1*, i351-358.
21. Camacho C.; Coulouris G.; Avagyan V.; Ma N.; Papadopoulos J.; Bealer K.; Madden T. L. BLAST+: architecture and applications. *BMC Bioinformatics* **2009**, *10*, 421.
22. She R.; Chu J. S.; Wang K.; Pei J.; Chen N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res* **2009**, *19*, 143-149.
23. Slater G. S.; Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **2005**, *6*, 31.
24. Stanke M.; Keller O.; Gunduz I.; Hayes A.; Waack S.; Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **2006**, *34*, W435-439.
25. Ghosh S.; Chan C. K. Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Methods Mol Biol* **2016**, *1374*, 339-361.
26. Li L.; Stoeckert C. J., Jr.; Roos D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **2003**, *13*, 2178-2189.
27. Tamura K.; Battistuzzi F. U.; Billings-Ross P.; Murillo O.; Filipowski A.; Kumar S. Estimating divergence times in large molecular phylogenies. *Proc Natl Acad Sci U S A* **2012**, *109*, 19333-19338.
28. Kumar S.; Stecher G.; Suleski M.; Hedges S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **2017**, *34*, 1812-1819.
29. Han M. V.; Thomas G. W.; Lugo-Martinez J.; Hahn M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* **2013**, *30*, 1987-1997.
30. Loytynoja A.; Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* **2005**, *102*, 10557-10562.
31. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **2007**, *24*, 1586-

1591.

32. Zhang J.; Nielsen R.; Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **2005**, *22*, 2472-2479.
33. Nielsen R.; Bustamante C.; Clark A. G.; Glanowski S.; Sackton T. B.; Hubisz M. J.; Fledel-Alon A.; Tanenbaum D. M.; Civello D.; White T. J.; J. J. S.; Adams M. D.; Cargill M. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **2005**, *3*, e170.
34. Li H.; Handsaker B.; Wysoker A.; Fennell T.; Ruan J.; Homer N.; Marth G.; Abecasis G.; Durbin R.; Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079.
35. Adzhubei I. A.; Schmidt S.; Peshkin L.; Ramensky V. E.; Gerasimova A.; Bork P.; Kondrashov A. S.; Sunyaev S. R. A method and server for predicting damaging missense mutations. *Nat Methods* **2010**, *7*, 248-249.
36. Choi Y.; Sims G. E.; Murphy S.; Miller J. R.; Chan A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **2012**, *7*, e46688.
37. Huang da W.; Sherman B. T.; Lempicki R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009**, *4*, 44-57.
38. Benjamini Y.; Draï D.; Elmer G.; Kafkafi N.; Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **2001**, *125*, 279-284.
39. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **2006**, *22*, 2688-2690.
40. Li H.; Durbin R. Inference of human population history from individual whole-genome sequences. *Nature* **2011**, *475*, 493-496.
41. Prado-Martinez J.; Sudmant P. H.; Kidd J. M.; Li H.; Kelley J. L.; Lorente-Galdos B.; Veeramah K. R.; Woerner A. E.; O'Connor T. D.; Santpere G.; Cagan A.; Theunert C.; Casals F.; Laayouni H.; Munch K.; Hobolth A.; Halager A. E.; Malig M.; Hernandez-Rodriguez J.; Hernando-Herraez I.; Prufer K.; Pybus M.; Johnstone L.; Lachmann M.; Alkan C.; Twigg D.; Petit N.; Baker C.; Hormozdiari F.; Fernandez-Callejo M.; Dabad M.; Wilson M. L.; Stevison L.; Camprubi C.; Carvalho T.; Ruiz-Herrera A.; Vives L.; Mele M.; Abello T.; Kondova I.; Bontrop R. E.; Pusey A.; Lankester F.; Kiyang J. A.; Bergl R. A.; Lonsdorf E.; Myers S.; Ventura M.; Gagneux P.; Comas D.; Siegmund H.; Blanc J.; Agueda-Calpena L.; Gut M.; Fulton L.; Tishkoff S. A.; Mullikin J. C.; Wilson R. K.; Gut I. G.; Gonder M. K.; Ryder O. A.; Hahn B. H.; Navarro A.; Akey J. M.; Bertranpetit J.; Reich D.; Mailund T.; Schierup M. H.; Hvilsom C.; Andres A. M.; Wall J. D.; Bustamante C. D.; Hammer M. F.; Eichler E. E.; Marques-Bonet T. Great ape genetic diversity and population history. *Nature* **2013**, *499*, 471-475.
42. Kaeuffer R.; Pontier D.; Devillard S.; Perrin N. Effective size of two feral domestic cat

- populations (*Felis catus* L): effect of the mating system. *Mol Ecol* **2004**, *13*, 483-490.
43. Zdobnov E. M.; Apweiler R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **2001**, *17*, 847-848.
 44. DuBuc T. Q.; Ryan J. F.; Shinzato C.; Satoh N.; Martindale M. Q. Coral comparative genomics reveal expanded Hox cluster in the cnidarian-bilaterian ancestor. *Integr Comp Biol* **2012**, *52*, 835-841.
 45. Chourrout D.; Delsuc F.; Chourrout P.; Edvardsen R. B.; Rentzsch F.; Renfer E.; Jensen M. F.; Zhu B.; de Jong P.; Steele R. E.; Technau U. Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature* **2006**, *442*, 684-687.
 46. Chiori R.; Jager M.; Denker E.; Wincker P.; Da Silva C.; Le Guyader H.; Manuel M.; Queinnec E. Are Hox genes ancestrally involved in axial patterning? Evidence from the hydrozoan *Clytia hemisphaerica* (Cnidaria). *PLoS One* **2009**, *4*, e4231.
 47. Price M. N.; Dehal P. S.; Arkin A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **2010**, *5*, e9490.
 48. Kusserow A.; Pang K.; Sturm C.; Hroudá M.; Lentfer J.; Schmidt H. A.; Technau U.; von Haeseler A.; Hobmayer B.; Martindale M. Q.; Holstein T. W. Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* **2005**, *433*, 156-160.
 49. Lengfeld T.; Watanabe H.; Simakov O.; Lindgens D.; Gee L.; Law L.; Schmidt H. A.; Ozbek S.; Bode H.; Holstein T. W. Multiple Wnts are involved in Hydra organizer formation and regeneration. *Dev Biol* **2009**, *330*, 186-199.
 50. Capella-Gutierrez S.; Silla-Martinez J. M.; Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **2009**, *25*, 1972-1973.
 51. Kurtz S.; Narechania A.; Stein J. C.; Ware D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **2008**, *9*, 517.
 52. Ellegren H.; Galtier N. Determinants of genetic diversity. *Nature Reviews Genetics* **2016**, *17*, 422-433.
 53. Putnam N. H.; Srivastava M.; Hellsten U.; Dirks B.; Chapman J.; Salamov A.; Terry A.; Shapiro H.; Lindquist E.; Kapitonov V. V.; Jurka J.; Genikhovich G.; Grigoriev I. V.; Lucas S. M.; Steele R. E.; Finnerty J. R.; Technau U.; Martindale M. Q.; Rokhsar D. S. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **2007**, *317*, 86-94.
 54. Chapman J. A.; Kirkness E. F.; Simakov O.; Hampson S. E.; Mitros T.; Weinmaier T.; Rattei T.; Balasubramanian P. G.; Borman J.; Busam D.; Disbennett K.; Pfannkoch C.; Sumin N.; Sutton G. G.; Viswanathan L. D.; Walenz B.; Goodstein D. M.; Hellsten U.; Kawashima T.; Prochnik S. E.; Putnam N. H.; Shu S.; Blumberg B.; Dana C. E.; Gee L.; Kibler D. F.; Law L.; Lindgens D.; Martinez D. E.; Peng J.; Wigge P. A.; Bertulat B.; Guder C.; Nakamura Y.; Ozbek S.; Watanabe H.;

- Khalturin K.; Hemmrich G.; Franke A.; Augustin R.; Fraune S.; Hayakawa E.; Hayakawa S.; Hirose M.; Hwang J. S.; Ikeo K.; Nishimiya-Fujisawa C.; Ogura A.; Takahashi T.; Steinmetz P. R.; Zhang X.; Aufschnaiter R.; Eder M. K.; Gorny A. K.; Salvenmoser W.; Heimberg A. M.; Wheeler B. M.; Peterson K. J.; Bottger A.; Tischler P.; Wolf A.; Gojobori T.; Remington K. A.; Strausberg R. L.; Venter J. C.; Technau U.; Hobmayer B.; Bosch T. C.; Holstein T. W.; Fujisawa T.; Bode H. R.; David C. N.; Rokhsar D. S.; Steele R. E. The dynamic genome of Hydra. *Nature* **2010**, *464*, 592-596.
55. Shinzato C.; Shoguchi E.; Kawashima T.; Hamada M.; Hisata K.; Tanaka M.; Fujie M.; Fujiwara M.; Koyanagi R.; Ikuta T. Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature* **2011**, *476*, 320.
56. Chen Y. C.; Liu T.; Yu C. H.; Chiang T. Y.; Hwang C. C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* **2013**, *8*, e62856.
57. van Dijk E. L.; Auger H.; Jaszczyszyn Y.; Thermes C. Ten years of next-generation sequencing technology. *Trends Genet* **2014**, *30*, 418-426.
58. Ashton P. M.; Nair S.; Dallman T.; Rubino S.; Rabsch W.; Mwaigwisya S.; Wain J.; O'Grady J. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* **2015**, *33*, 296-300.
59. Koren S.; Phillippy A. M. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol* **2015**, *23*, 110-120.
60. Koren S.; Harhay G. P.; Smith T. P.; Bono J. L.; Harhay D. M.; McVey S. D.; Radune D.; Bergman N. H.; Phillippy A. M. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol* **2013**, *14*, R101.
61. Jain M.; Fiddes I. T.; Miga K. H.; Olsen H. E.; Paten B.; Akeson M. Improved data analysis for the MinION nanopore sequencer. *Nat Methods* **2015**, *12*, 351-356.
62. Compeau P. E.; Pevzner P. A.; Tesler G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* **2011**, *29*, 987-991.
63. Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinform* **2009**, *10*, 354-366.
64. Miller J. R.; Koren S.; Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* **2010**, *95*, 315-327.
65. Bankevich A.; Nurk S.; Antipov D.; Gurevich A. A.; Dvorkin M.; Kulikov A. S.; Lesin V. M.; Nikolenko S. I.; Pham S.; Prjibelski A. D.; Pyshkin A. V.; Sirotkin A. V.; Vyahhi N.; Tesler G.; Alekseyev M. A.; Pevzner P. A. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **2012**, *19*, 455-477.
66. Butler J.; MacCallum I.; Kleber M.; Shlyakhter I. A.; Belmonte M. K.; Lander E. S.; Nusbaum C.;

- Jaffe D. B. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* **2008**, *18*, 810-820.
67. Silva G. G.; Dutilh B. E.; Matthews T. D.; Elkins K.; Schmieder R.; Dinsdale E. A.; Edwards R. A. Combining de novo and reference-guided assembly with scaffold_builder. *Source Code Biol Med* **2013**, *8*, 23.
68. Denisov G.; Walenz B.; Halpern A. L.; Miller J.; Axelrod N.; Levy S.; Sutton G. Consensus generation and variant detection by Celera Assembler. *Bioinformatics* **2008**, *24*, 1035-1040.
69. Koren S.; Walenz B. P.; Berlin K.; Miller J. R.; Bergman N. H.; Phillippy A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **2017**, *27*, 722-736.
70. Kamath G. M.; Shomorony I.; Xia F.; Courtade T. A.; Tse D. N. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Res* **2017**, *27*, 747-756.
71. Kajitani R.; Toshimoto K.; Noguchi H.; Toyoda A.; Ogura Y.; Okuno M.; Yabana M.; Harada M.; Nagayasu E.; Maruyama H.; Kohara Y.; Fujiyama A.; Hayashi T.; Itoh T. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* **2014**, *24*, 1384-1395.
72. Maccallum I.; Przybylski D.; Gnerre S.; Burton J.; Shlyakhter I.; Gnirke A.; Malek J.; McKernan K.; Ranade S.; Shea T. P.; Williams L.; Young S.; Nusbaum C.; Jaffe D. B. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol* **2009**, *10*, R103.
73. Simpson J. T.; Wong K.; Jackman S. D.; Schein J. E.; Jones S. J.; Birol I. ABySS: a parallel assembler for short read sequence data. *Genome Res* **2009**, *19*, 1117-1123.
74. Zerbino D. R.; Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **2008**, *18*, 821-829.
75. Ye C.; Ma Z. S.; Cannon C. H.; Pop M.; Yu D. W. Exploiting sparseness in de novo genome assembly. *BMC Bioinformatics* **2012**, *13 Suppl 6*, S1.
76. Simpson J. T.; Durbin R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* **2012**, *22*, 549-556.
77. Zimin A. V.; Marcais G.; Puiu D.; Roberts M.; Salzberg S. L.; Yorke J. A. The MaSuRCA genome assembler. *Bioinformatics* **2013**, *29*, 2669-2677.
78. Earl D.; Bradnam K.; St John J.; Darling A.; Lin D.; Fass J.; Yu H. O.; Buffalo V.; Zerbino D. R.; Diekhans M.; Nguyen N.; Ariyaratne P. N.; Sung W. K.; Ning Z.; Haimel M.; Simpson J. T.; Fonseca N. A.; Birol I.; Docking T. R.; Ho I. Y.; Rokhsar D. S.; Chikhi R.; Lavenier D.; Chapuis G.; Naquin D.; Maillet N.; Schatz M. C.; Kelley D. R.; Phillippy A. M.; Koren S.; Yang S. P.; Wu W.; Chou W. C.; Srivastava A.; Shaw T. I.; Ruby J. G.; Skewes-Cox P.; Betegon M.; Dimon M. T.;

- Solovyev V.; Seledtsov I.; Kosarev P.; Vorobyev D.; Ramirez-Gonzalez R.; Leggett R.; MacLean D.; Xia F.; Luo R.; Li Z.; Xie Y.; Liu B.; Gnerre S.; MacCallum I.; Przybylski D.; Ribeiro F. J.; Yin S.; Sharpe T.; Hall G.; Kersey P. J.; Durbin R.; Jackman S. D.; Chapman J. A.; Huang X.; DeRisi J. L.; Caccamo M.; Li Y.; Jaffe D. B.; Green R. E.; Haussler D.; Korf I.; Paten B. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* **2011**, *21*, 2224-2241.
79. Bradnam K. R.; Fass J. N.; Alexandrov A.; Baranay P.; Bechner M.; Birol I.; Boisvert S.; Chapman J. A.; Chapuis G.; Chikhi R.; Chitsaz H.; Chou W. C.; Corbeil J.; Del Fabbro C.; Docking T. R.; Durbin R.; Earl D.; Emrich S.; Fedotov P.; Fonseca N. A.; Ganapathy G.; Gibbs R. A.; Gnerre S.; Godzaridis E.; Goldstein S.; Haimel M.; Hall G.; Haussler D.; Hiatt J. B.; Ho I. Y.; Howard J.; Hunt M.; Jackman S. D.; Jaffe D. B.; Jarvis E. D.; Jiang H.; Kazakov S.; Kersey P. J.; Kitzman J. O.; Knight J. R.; Koren S.; Lam T. W.; Lavenier D.; Laviolette F.; Li Y.; Li Z.; Liu B.; Liu Y.; Luo R.; Maccallum I.; Macmanes M. D.; Maillet N.; Melnikov S.; Naquin D.; Ning Z.; Otto T. D.; Paten B.; Paulo O. S.; Phillippy A. M.; Pina-Martins F.; Place M.; Przybylski D.; Qin X.; Qu C.; Ribeiro F. J.; Richards S.; Rokhsar D. S.; Ruby J. G.; Scalabrin S.; Schatz M. C.; Schwartz D. C.; Sergushichev A.; Sharpe T.; Shaw T. I.; Shendure J.; Shi Y.; Simpson J. T.; Song H.; Tsarev F.; Vezzi F.; Vicedomini R.; Vieira B. M.; Wang J.; Worley K. C.; Yin S.; Yiu S. M.; Yuan J.; Zhang G.; Zhang H.; Zhou S.; Korf I. F. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2013**, *2*, 10.
80. Salmela L.; Walve R.; Rivals E.; Ukkonen E. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* **2017**, *33*, 799-806.
81. McCoy R. C.; Taylor R. W.; Blauwkamp T. A.; Kelley J. L.; Kertesz M.; Pushkarev D.; Petrov D. A.; Fiston-Lavier A. S. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **2014**, *9*, e106689.
82. Sleator R. D. An overview of the current status of eukaryote gene prediction strategies. *Gene* **2010**, *461*, 1-4.
83. Pontius J. U.; Mullikin J. C.; Smith D. R.; Agencourt Sequencing T.; Lindblad-Toh K.; Gnerre S.; Clamp M.; Chang J.; Stephens R.; Neelam B.; Volfovsky N.; Schaffer A. A.; Agarwala R.; Narfstrom K.; Murphy W. J.; Giger U.; Roca A. L.; Antunes A.; Menotti-Raymond M.; Yuhki N.; Pecon-Slattery J.; Johnson W. E.; Bourque G.; Tesler G.; Program N. C. S.; O'Brien S. J. Initial sequence and comparative analysis of the cat genome. *Genome Res* **2007**, *17*, 1675-1689.
84. Montague M. J.; Li G.; Gandolfi B.; Khan R.; Aken B. L.; Searle S. M.; Minx P.; Hillier L. W.; Koboldt D. C.; Davis B. W.; Driscoll C. A.; Barr C. S.; Blackstone K.; Quilez J.; Lorente-Galdos B.; Marques-Bonet T.; Alkan C.; Thomas G. W.; Hahn M. W.; Menotti-Raymond M.; O'Brien S. J.; Wilson R. K.; Lyons L. A.; Murphy W. J.; Warren W. C. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc Natl Acad Sci*

USA 2014, III, 17230-17235.

85. Dobrynin P.; Liu S.; Tamazian G.; Xiong Z.; Yurchenko A. A.; Krasheninnikova K.; Kliver S.; Schmidt-Kuntzel A.; Koepfli K. P.; Johnson W.; Kuderna L. F.; Garcia-Perez R.; Manuel M.; Godinez R.; Komissarov A.; Makunin A.; Brukhin V.; Qiu W.; Zhou L.; Li F.; Yi J.; Driscoll C.; Antunes A.; Oleksyk T. K.; Eizirik E.; Perelman P.; Roelke M.; Wildt D.; Diekhans M.; Marques-Bonet T.; Marker L.; Bhak J.; Wang J.; Zhang G.; O'Brien S. J. Genomic legacy of the African cheetah, *Acinonyx jubatus*. *Genome Biol* **2015**, *16*, 277.
86. Bejerano G.; Pheasant M.; Makunin I.; Stephen S.; Kent W. J.; Mattick J. S.; Haussler D. Ultraconserved elements in the human genome. *Science* **2004**, *304*, 1321-1325.
87. Woolfe A.; Goodson M.; Goode D. K.; Snell P.; McEwen G. K.; Vavouri T.; Smith S. F.; North P.; Callaway H.; Kelly K.; Walter K.; Abnizova I.; Gilks W.; Edwards Y. J.; Cooke J. E.; Elgar G. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **2005**, *3*, e7.
88. Qiu Q.; Zhang G.; Ma T.; Qian W.; Wang J.; Ye Z.; Cao C.; Hu Q.; Kim J.; Larkin D. M.; Auvil L.; Capitanu B.; Ma J.; Lewin H. A.; Qian X.; Lang Y.; Zhou R.; Wang L.; Wang K.; Xia J.; Liao S.; Pan S.; Lu X.; Hou H.; Wang Y.; Zang X.; Yin Y.; Ma H.; Zhang J.; Wang Z.; Zhang Y.; Zhang D.; Yonezawa T.; Hasegawa M.; Zhong Y.; Liu W.; Zhang Y.; Huang Z.; Zhang S.; Long R.; Yang H.; Wang J.; Lenstra J. A.; Cooper D. N.; Wu Y.; Wang J.; Shi P.; Wang J.; Liu J. The yak genome and adaptation to life at high altitude. *Nat Genet* **2012**, *44*, 946-949.
89. Locke D. P.; Hillier L. W.; Warren W. C.; Worley K. C.; Nazareth L. V.; Muzny D. M.; Yang S. P.; Wang Z.; Chinwalla A. T.; Minx P.; Mitreva M.; Cook L.; Delehaunty K. D.; Fronick C.; Schmidt H.; Fulton L. A.; Fulton R. S.; Nelson J. O.; Magrini V.; Pohl C.; Graves T. A.; Markovic C.; Cree A.; Dinh H. H.; Hume J.; Kovar C. L.; Fowler G. R.; Lunter G.; Meader S.; Heger A.; Ponting C. P.; Marques-Bonet T.; Alkan C.; Chen L.; Cheng Z.; Kidd J. M.; Eichler E. E.; White S.; Searle S.; Vilella A. J.; Chen Y.; Flicek P.; Ma J.; Raney B.; Suh B.; Burhans R.; Herrero J.; Haussler D.; Faria R.; Fernando O.; Darre F.; Farre D.; Gazave E.; Oliva M.; Navarro A.; Roberto R.; Capozzi O.; Archidiacono N.; Della Valle G.; Purgato S.; Rocchi M.; Konkel M. K.; Walker J. A.; Ullmer B.; Batzer M. A.; Smit A. F.; Hubley R.; Casola C.; Schrider D. R.; Hahn M. W.; Quesada V.; Puente X. S.; Ordonez G. R.; Lopez-Otin C.; Vinar T.; Brejova B.; Ratan A.; Harris R. S.; Miller W.; Kosiol C.; Lawson H. A.; Taliwal V.; Martins A. L.; Siepel A.; Roychoudhury A.; Ma X.; Degenhardt J.; Bustamante C. D.; Gutenkunst R. N.; Mailund T.; Dutheil J. Y.; Hobolth A.; Schierup M. H.; Ryder O. A.; Yoshinaga Y.; de Jong P. J.; Weinstock G. M.; Rogers J.; Mardis E. R.; Gibbs R. A.; Wilson R. K. Comparative and demographic analysis of orang-utan genomes. *Nature* **2011**, *469*, 529-533.
90. Li R.; Fan W.; Tian G.; Zhu H.; He L.; Cai J.; Huang Q.; Cai Q.; Li B.; Bai Y.; Zhang Z.; Zhang Y.; Wang W.; Li J.; Wei F.; Li H.; Jian M.; Li J.; Zhang Z.; Nielsen R.; Li D.; Gu W.; Yang Z.; Xuan Z.;

- Ryder O. A.; Leung F. C.; Zhou Y.; Cao J.; Sun X.; Fu Y.; Fang X.; Guo X.; Wang B.; Hou R.; Shen F.; Mu B.; Ni P.; Lin R.; Qian W.; Wang G.; Yu C.; Nie W.; Wang J.; Wu Z.; Liang H.; Min J.; Wu Q.; Cheng S.; Ruan J.; Wang M.; Shi Z.; Wen M.; Liu B.; Ren X.; Zheng H.; Dong D.; Cook K.; Shan G.; Zhang H.; Kosiol C.; Xie X.; Lu Z.; Zheng H.; Li Y.; Steiner C. C.; Lam T. T.; Lin S.; Zhang Q.; Li G.; Tian J.; Gong T.; Liu H.; Zhang D.; Fang L.; Ye C.; Zhang J.; Hu W.; Xu A.; Ren Y.; Zhang G.; Bruford M. W.; Li Q.; Ma L.; Guo Y.; An N.; Hu Y.; Zheng Y.; Shi Y.; Li Z.; Liu Q.; Chen Y.; Zhao J.; Qu N.; Zhao S.; Tian F.; Wang X.; Wang H.; Xu L.; Liu X.; Vinar T.; Wang Y.; Lam T. W.; Yiu S. M.; Liu S.; Zhang H.; Li D.; Huang Y.; Wang X.; Yang G.; Jiang Z.; Wang J.; Qin N.; Li L.; Li J.; Bolund L.; Kristiansen K.; Wong G. K.; Olson M.; Zhang X.; Li S.; Yang H.; Wang J.; Wang J. The sequence and de novo assembly of the giant panda genome. *Nature* **2010**, *463*, 311-317.
91. Irizarry K. J.; Malladi S. B.; Gao X.; Mitsouras K.; Melendez L.; Burriss P. A.; Brockman J. A.; Al-Murrani S. W. Sequencing and comparative genomic analysis of 1227 *Felis catus* cDNA sequences enriched for developmental, clinical and nutritional phenotypes. *BMC Genomics* **2012**, *13*, 31.
92. Yang Z.; Bielawski J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **2000**, *15*, 496-503.
93. Ferguson L. R. Meat and cancer. *Meat Sci* **2010**, *84*, 308-313.
94. Bastide N. M.; Pierre F. H.; Corpet D. E. Heme iron from meat and risk of colorectal cancer: a meta-analysis and a review of the mechanisms involved. *Cancer Prev Res (Phila)* **2011**, *4*, 177-184.
95. Oostindjer M.; Alexander J.; Amdam G. V.; Andersen G.; Bryan N. S.; Chen D.; Corpet D. E.; De Smet S.; Dragsted L. O.; Haug A.; Karlsson A. H.; Kleter G.; de Kok T. M.; Kulseng B.; Milkowski A. L.; Martin R. J.; Pajari A. M.; Paulsen J. E.; Pickova J.; Rudi K.; Sodring M.; Weed D. L.; Egelanddal B. The role of red and processed meat in colorectal cancer development: a perspective. *Meat Sci* **2014**, *97*, 583-596.
96. Schermerhorn T. Normal glucose metabolism in carnivores overlaps with diabetes pathology in non-carnivores. *Front Endocrinol (Lausanne)* **2013**, *4*, 188.
97. Siepel A.; Bejerano G.; Pedersen J. S.; Hinrichs A. S.; Hou M.; Rosenbloom K.; Clawson H.; Spieth J.; Hillier L. W.; Richards S.; Weinstock G. M.; Wilson R. K.; Gibbs R. A.; Kent W. J.; Miller W.; Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **2005**, *15*, 1034-1050.
98. Oleksyk T. K.; Smith M. W.; O'Brien S. J. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci* **2010**, *365*, 185-205.
99. Johnson W. E.; Eizirik E.; Pecon-Slattery J.; Murphy W. J.; Antunes A.; Teeling E.; O'Brien S. J. The late Miocene radiation of modern Felidae: a genetic assessment. *Science* **2006**, *311*, 73-77.
100. O'Brien S. J.; Johnson W. E. The evolution of cats. Genomic paw prints in the DNA of the

- world's wild cats have clarified the cat family tree and uncovered several remarkable migrations in their past. *Sci Am* **2007**, *297*, 68-75.
101. Camazine S. Olfactory aposematism : Association of food toxicity with naturally occurring odor. *J Chem Ecol* **1985**, *11*, 1289-1295.
102. Van Valkenburgh B. Major patterns in the history of carnivorous mammals. *Annual Review of Earth and Planetary Sciences* **1999**, *27*, 463-493.
103. Van Valkenburgh B.; Wang X.; Damuth J. Cope's rule, hypercarnivory, and extinction in North American canids. *Science* **2004**, *306*, 101-104.
104. Forrest J. L.; Wikramanayake E.; Shrestha R.; Arendran G.; Gyeltshen K.; Maheshwari A.; Mazumdar S.; Naidoo R.; Thapa G. J.; Thapa K. Conservation and climate change: Assessing the vulnerability of snow leopard habitat to treeline shift in the Himalaya. *Biological Conservation* **2012**, *150*, 129-135.
105. Luo S. J.; Zhang Y.; Johnson W. E.; Miao L.; Martelli P.; Antunes A.; Smith J. L.; O'Brien S. J. Sympatric Asian felid phylogeography reveals a major Indochinese-Sundaic divergence. *Mol Ecol* **2014**, *23*, 2072-2092.
106. Bond J. S.; Beynon R. J. The astacin family of metalloendopeptidases. *Protein Sci* **1995**, *4*, 1247-1261.
107. Bork P.; Beckmann G. The CUB domain. A widespread module in developmentally regulated proteins. *J Mol Biol* **1993**, *231*, 539-545.
108. Rachamim T.; Morgenstern D.; Aharonovich D.; Brekhman V.; Lotan T.; Sher D. The dynamically evolving nematocyst content of an anthozoan, a scyphozoan, and a hydrozoan. *Mol Biol Evol* **2015**, *32*, 740-753.
109. Brekhman V.; Malik A.; Haas B.; Sher N.; Lotan T. Transcriptome profiling of the dynamic life cycle of the scyphozoan jellyfish *Aurelia aurita*. *BMC Genomics* **2015**, *16*, 74.
110. Holland P. W. Evolution of homeobox genes. *Wiley Interdiscip Rev Dev Biol* **2013**, *2*, 31-45.
111. Kamm K.; Schierwater B.; Jakob W.; Dellaporta S. L.; Miller D. J. Axial patterning and diversification in the cnidaria predate the Hox system. *Curr Biol* **2006**, *16*, 920-926.
112. Finnerty J. R.; Pang K.; Burton P.; Paulson D.; Martindale M. Q. Origins of bilateral symmetry: Hox and dpp expression in a sea anemone. *Science* **2004**, *304*, 1335-1337.
113. Nusse R. An ancient cluster of Wnt paralogues. *Trends Genet* **2001**, *17*, 443.
114. Denker A.; de Laat W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev* **2016**, *30*, 1357-1382.
115. de Wit E.; de Laat W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **2012**, *26*, 11-24.
116. Dong Y.; Xie M.; Jiang Y.; Xiao N.; Du X.; Zhang W.; Tossier-Klopp G.; Wang J.; Yang S.; Liang

J.; Chen W.; Chen J.; Zeng P.; Hou Y.; Bian C.; Pan S.; Li Y.; Liu X.; Wang W.; Servin B.; Sayre B.; Zhu B.; Sweeney D.; Moore R.; Nie W.; Shen Y.; Zhao R.; Zhang G.; Li J.; Faraut T.; Womack J.; Zhang Y.; Kijas J.; Cockett N.; Xu X.; Zhao S.; Wang J.; Wang W. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat Biotechnol* **2013**, *31*, 135-141.

Acknowledgements

This work was supported by the National Institute of Biological Resources of Korea in-house program (NIBR201503101, NIBR201603104). This work was also supported by the 2015 Research fund (1.150014.01) of Ulsan National Institute of Science & Technology (UNIST). This work was also supported by the Genome Korea Project in Ulsan Research Funds (1.180024.01 and 1.180017.01) of Ulsan National Institute of Science & Technology (UNIST). This work was also supported by a grant from the Marine Biotechnology Program (20170305, Development of Biomedical materials based on marine proteins) and the Collaborative Genome Program (20140428) funded by the Ministry of Oceans and Fisheries, Korea. This work was also supported by the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (NRF-2017M3C9A6047623 and NRF-2017R1A2B2012541). Korea Institute of Science and Technology Information (KISTI) provided us with Korea Research Environment Open NETwork (KREONET), which is the internet connection service for efficient information and data transfer.

I express my sincere gratitude to my advisor, professor Jong Bhak, for his invaluable advice and continuous support. My committee members, Prof. Cheol-Min Ghim, Prof. Dougu Nam, Prof. Semin Lee, and Prof. Seung woo Cho are thanked for their critical comments and thoughtful suggestions.

I would like to thank all my colleagues, for their unwavering support and encouragement. I have always been lucky to have good colleagues to work with.

Finally, I would like to express my deeply thanks to my wife, son, and parent for their love, understanding and forbearance.

Appendix

The leopard reference genome

Kim *et al. Genome Biology* (2016) 17:211
DOI 10.1186/s13059-016-1071-4

Genome Biology

RESEARCH

Open Access



Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly

Soonok Kim^{1†}, Yun Sung Cho^{2,3,4†}, Hak-Min Kim^{2,3†}, Oksung Chung⁴, Hyunho Kim⁵, Sungwoong Jho⁴, Hong Seomun⁶, Jeongho Kim⁷, Woo Young Bang¹, Changmu Kim¹, Junghwa An⁶, Chang Hwan Bae¹, Youngjune Bhak², Sungwon Jeon^{2,3}, Hyejun Yoon^{2,3}, Yumi Kim², JeHoon Jun^{4,5}, HyeJin Lee^{4,5}, Suan Cho^{4,5}, Olga Uphyrkina⁸, Aleksey Kostyria⁸, John Goodrich⁹, Dale Miquelle^{10,11}, Melody Roelke¹², John Lewis¹³, Andrey Yurchenko¹⁴, Anton Bankevich¹⁵, Juok Cho¹⁶, Semin Lee^{2,3,17}, Jeremy S. Edwards¹⁸, Jessica A. Weber¹⁹, Jo Cook²⁰, Sangsoo Kim²¹, Hang Lee²², Andrea Manica²³, Ilbeum Lee²⁴, Stephen J. O'Brien^{14,25*}, Jong Bhak^{2,3,4,5*} and Joo-Hong Yeo^{1*}

Abstract

Background: There are three main dietary groups in mammals: carnivores, omnivores, and herbivores. Currently, there is limited comparative genomics insight into the evolution of dietary specializations in mammals. Due to recent advances in sequencing technologies, we were able to perform in-depth whole genome analyses of representatives of these three dietary groups.

Results: We investigated the evolution of carnivory by comparing 18 representative genomes from across Mammalia with carnivorous, omnivorous, and herbivorous dietary specializations, focusing on Felidae (domestic cat, tiger, lion, cheetah, and leopard), Hominidae, and Bovidae genomes. We generated a new high-quality leopard genome assembly, as well as two wild Amur leopard whole genomes. In addition to a clear contraction in gene families for starch and sucrose metabolism, the carnivore genomes showed evidence of shared evolutionary adaptations in genes associated with diet, muscle strength, agility, and other traits responsible for successful hunting and meat consumption. Additionally, an analysis of highly conserved regions at the family level revealed molecular signatures of dietary adaptation in each of Felidae, Hominidae, and Bovidae. However, unlike carnivores, omnivores and herbivores showed fewer shared adaptive signatures, indicating that carnivores are under strong selective pressure related to diet. Finally, felids showed recent reductions in genetic diversity associated with decreased population sizes, which may be due to the inflexible nature of their strict diet, highlighting their vulnerability and critical conservation status.

Conclusions: Our study provides a large-scale family level comparative genomic analysis to address genomic changes associated with dietary specialization. Our genomic analyses also provide useful resources for diet-related genetic and health research.

Keywords: Carnivorous diet, Evolutionary adaptation, Leopard, Felidae, *De novo* assembly, Comparative genomics

* Correspondence: lgdchief@gmail.com; jongbhak@genomics.org;
y1208@korea.kr

†Equal contributors

¹⁴Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg 199004, Russia

²The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

¹Biological and Genetic Resources Assessment Division, National Institute of Biological Resources, Incheon 22689, Republic of Korea

Full list of author information is available at the end of the article



© 2016 The Author(s). **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

RESEARCH ARTICLE

Open Access

The genome of the giant Nomura's jellyfish sheds light on the early evolution of active predation



Hak-Min Kim^{1,2†}, Jessica A. Weber^{3,4†}, Nayoung Lee^{5†}, Seung Gu Park¹, Yun Sung Cho^{1,2,6}, Youngjune Bhak^{1,2}, Nayun Lee⁵, Yeonsu Jeon^{1,2}, Sungwon Jeon^{1,2}, Victor Luria⁷, Amir Karger⁸, Marc W. Kirschner⁷, Ye Jin Jo⁵, Seonock Woo^{9,10}, Kyoungsoo Shin¹¹, Oksung Chung^{6,12}, Jae-Chun Ryu¹³, Hyung-Soon Yim¹⁰, Jung-Hyun Lee¹⁰, Jeremy S. Edwards¹⁴, Andrea Manica¹⁵, Jong Bhak^{1,2,6,12*} and Seungshic Yum^{5,9*}

Abstract

Background: Unique among cnidarians, jellyfish have remarkable morphological and biochemical innovations that allow them to actively hunt in the water column and were some of the first animals to become free-swimming. The class Scyphozoa, or true jellyfish, are characterized by a predominant medusa life-stage consisting of a bell and venomous tentacles used for hunting and defense, as well as using pulsed jet propulsion for mobility. Here, we present the genome of the giant Nomura's jellyfish (*Nemopilema nomurai*) to understand the genetic basis of these key innovations.

Results: We sequenced the genome and transcriptomes of the bell and tentacles of the giant Nomura's jellyfish as well as transcriptomes across tissues and developmental stages of the *Sanderia malayensis* jellyfish. Analyses of the *Nemopilema* and other cnidarian genomes revealed adaptations associated with swimming, marked by codon bias in muscle contraction and expansion of neurotransmitter genes, along with expanded Myosin type II family and venom domains, possibly contributing to jellyfish mobility and active predation. We also identified gene family expansions of *Wnt* and posterior *Hox* genes and discovered the important role of retinoic acid signaling in this ancient lineage of metazoans, which together may be related to the unique jellyfish body plan (medusa formation).

Conclusions: Taken together, the *Nemopilema* jellyfish genome and transcriptomes genetically confirm their unique morphological and physiological traits, which may have contributed to the success of jellyfish as early multi-cellular predators.

Keywords: Jellyfish mobility, Medusa structure formation, Scyphozoa, de novo genome assembly

Background

Cnidarians, including jellyfish and their predominantly sessile relatives the coral, sea anemone, and hydra, first appeared in the Precambrian Era and are now key members of aquatic ecosystems worldwide (Fig. 1a) [1]. Between 500 and 700 million years ago, jellyfish developed novel physiological traits that allowed them to become

one of the first free-swimming predators. The life cycle of the jellyfish includes a small polypoid, sessile stage which reproduces asexually to form the mobile medusa form that can reproduce both sexually and asexually (Fig. 1c) [2]. The class Scyphozoa, or true jellyfish, are characterized by a predominant medusa life-stage consisting of a bell and venomous tentacles used for hunting and defense [3]. Jellyfish medusae feature a radially symmetric body structure, powered by readily identifiable cell types such as motor neurons and striated muscles that expand and contract to create the most energy-efficient swimming method in the animal kingdom [4, 5]. Over 95% water, jellyfish are osmoconformers that use ion gradients to deliver solutes to cells

* Correspondence: jongbhak@gmail.com; syum@kio.st.ac.kr

[†]Hak-Min Kim, Jessica A. Weber and Nayoung Lee contributed equally to this work.

¹Korean Genomics Industrialization Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

⁵Ecological Risk Research Division, Korea Institute of Ocean Science and Technology (KIOST), Geoje 53201, Republic of Korea

Full list of author information is available at the end of the article



© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

The Draft Genome of an Octocoral, *Dendronephthya gigantea*

Yeonsu Jeon^{1,2,†}, Seung Gu Park^{1,†}, Nayun Lee^{3,†}, Jessica A. Weber^{4,5}, Hui-Su Kim¹, Sung-Jin Hwang⁶, Seonock Woo⁷, Hak-Min Kim^{1,2}, Youngjune Bhak^{1,2}, Sungwon Jeon^{1,2}, Nayoung Lee³, Yejin Jo³, Asta Blazyte¹, Taewoo Ryu⁸, Yun Sung Cho^{1,2,9}, Hyunho Kim¹⁰, Jung-Hyun Lee⁷, Hyung-Soon Yim⁷, Jong Bhak^{1,2,9,10,*}, and Seungshic Yum^{3,11,*}

¹Korean Genomics Industrialization and Commercialization Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

²Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

³Ecological Risk Research Division, Korea Institute of Ocean Science and Technology (KIOST), Geoje, Republic of Korea

⁴Department of Genetics, Harvard Medical School, Boston, Massachusetts

⁵Department of Biology, University of New Mexico

⁶Department of Life Science, Woosuk University, Republic of Korea

⁷Marine Biotechnology Research Center, Korea Institute of Ocean Science and Technology (KIOST), Busan, Republic of Korea

⁸APEC Climate Center, Busan, South Korea

⁹Clinomics Inc., Ulsan, Republic of Korea

¹⁰Personal Genomics Institute, Genome Research Foundation, Cheongju, Republic of Korea

¹¹Faculty of Marine Environmental Science, University of Science and Technology (UST), Geoje, Republic of Korea

[†]These authors contributed equally to this work.

^{*}These authors jointly supervised this work.

*Corresponding authors: E-mails: jongbhak@genomics.org; syum@kiost.ac.kr.

Accepted: February 26, 2019

Data deposition: The octocoral whole genome and transcriptome project has been deposited at DDBJ/ENA/GenBank under the accession PRJNA507923 and PRJNA507943. DNA and RNA sequencing reads have been uploaded to the NCBI Read Archive under the accession (SRR8293699, and SRR8293698 and SRR8293935, and SRR8293936), respectively. The genome assembly has been deposited at DDBJ/ENA/GenBank under the accession RSEI01000000.

Abstract

Coral reefs composed of stony corals are threatened by global marine environmental changes. However, soft coral communities of octocorallian species, appear more resilient. The genomes of several cnidarians species have been published, including from stony corals, sea anemones, and hydra. To fill the phylogenetic gap for octocoral species of cnidarians, we sequenced the octocoral, *Dendronephthya gigantea*, a nonsymbiotic soft coral, commonly known as the carnation coral. The *D. gigantea* genome size is ~276 Mb. A high-quality genome assembly was constructed from PacBio long reads (29.85 Gb with 108× coverage) and Illumina short paired-end reads (35.54 Gb with 128× coverage) resulting in the highest N50 value (1.4 Mb) reported thus far among cnidarian genomes. About 12% of the genome is repetitive elements and contained 28,879 predicted protein-coding genes. This gene set is composed of 94% complete BUSCO ortholog benchmark genes, which is the second highest value among the cnidarians, indicating high quality. Based on molecular phylogenetic analysis, octocoral and hexacoral divergence times were estimated at 544 MYA. There is a clear difference in *Hox* gene composition between these species: unlike hexacorals, the Antp superclass *Evx* gene was absent in *D. gigantea*. Here, we present the first genome assembly of a nonsymbiotic octocoral, *D. gigantea* to aid in the comparative genomic analysis of cnidarians, including stony and soft corals, both symbiotic and nonsymbiotic. The *D. gigantea* genome may also provide clues to mechanisms of differential coping between the soft and stony corals in response to scenarios of global warming.

Key words: soft coral, genome, octocoral, nonsymbiotic coral, cnidarian, *Dendronephthya gigantea*.

© The Author(s) 2019. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

The *Galleria mellonella* Hologenome Supports Microbiota-Independent Metabolism of Long-Chain Hydrocarbon Beeswax

Hyun Gi Kong,^{1,4} Hyun Ho Kim,^{3,4} Joon-hui Chung,^{1,4} JeHoon Jun,³ Soohyun Lee,¹ Hak-Min Kim,² Sungwon Jeon,² Seung Gu Park,² Jong Bhak,^{2,3} and Choong-Min Ryu^{1,5,*}

¹Molecular Phytobacteriology Laboratory, Infection Disease Research Center, KRIBB, Daejeon 34141, South Korea

²Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea

³The Clinomics Institute, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea

⁴These authors contributed equally

⁵Lead Contact

*Correspondence: cmryu@kribb.re.kr

<https://doi.org/10.1016/j.celrep.2019.02.018>

SUMMARY

The greater wax moth, *Galleria mellonella*, degrades wax and plastic molecules. Despite much interest, the genetic basis of these hallmark traits remains poorly understood. Herein, we assembled high-quality genome and transcriptome data from *G. mellonella* to investigate long-chain hydrocarbon wax metabolism strategies. Specific carboxylesterase and lipase and fatty-acid-metabolism-related enzymes in the *G. mellonella* genome are transcriptionally regulated during feeding on beeswax. Strikingly, *G. mellonella* lacking intestinal microbiota successfully decomposes long-chain fatty acids following wax metabolism, although the intestinal microbiome performs a supplementary role in short-chain fatty acid degradation. Notably, final wax derivatives were detected by gas chromatography even in the absence of gut microbiota. Our findings provide insight into wax moth adaptation and may assist in the development of unique wax-degradation strategies with a similar metabolic approach for a plastic molecule polyethylene biodegradation using organisms without intestinal microbiota.

INTRODUCTION

In recent decades, plastics have been routinely released into the environment via sewage treatment plants, waste disposal, and aerial deposition, and global plastic production has expanded tremendously worldwide (Nowack and Bucheli, 2007). Plastic disposal is one of the biggest problems facing the environment, because vast amounts of synthetic plastic remain nondegradable (Nkwachukwu et al., 2013). Plastics are synthetic polymers composed of carbon, hydrogen, oxygen, and chloride that are derived from multiple sources, such as petroleum, coal, and natural gas. The most widely used plastics polymers are polyethylene (PE), polypropylene (PP), PE terephthalate (PET), polystyrene (PS), and polyvinyl chloride (PVC) (Wu et al., 2017). PE,

the most common petroleum-based plastic, is widely used in everyday life. However, the high durability and short usage time of PE is resulting in rapid accumulation in the environment, raising international interest (Ammala et al., 2011; Roy et al., 2011; Shah et al., 2008; Zettler et al., 2013).

The potential to decompose plastics in various environments has been studied in order to investigate biological degradation as a solution to accumulating plastics in the environment (Albertsson and Karlsson, 1988; Artham et al., 2009; Jones et al., 1974; Ohtake et al., 1998; Pegram and Andrady, 1989). Biodegradation of PE in the environment occurs mainly through the biological activity of microorganisms after thermal oxidation (Albertsson et al., 1987; Tokiwa et al., 2009). PE is decomposed into low-molecular-weight substances such as alkanes, alkenes, ketones, aldehydes, various alcohols, and fatty acids (Albertsson et al., 1987, 1998; Tokiwa et al., 2009). More than 90 genera of bacteria and fungi have been proposed to possess the ability to break down plastics (Mahdiyah and Mukti, 2013). However, many plastic components are recalcitrant to biodegradation by microorganisms, and the processing capacity is a generally very slow (Singh and Gupta, 2014). Metabolism of long-chain hydrocarbons is the most important step in the biodegradation of PE. This activity has not previously been reported in microorganisms. Interestingly, naturally occurring beeswax is a natural substance consisting of palmitoleate, long-chain aliphatic alcohols, and hydrocarbons. Similarly, PE is composed of a long-chain linear backbone of carbon atoms. The production of long-chain fatty acids and long-chain ethanol from beeswax is the most important process in long-chain hydrocarbon degradation. However, the associated genes and enzymes have not been studied in microorganisms.

Alternatively, the potential to metabolize long-chain hydrocarbons using insects has been studied extensively, because the enzymes and mechanisms mediating the biodegradation of long-chain hydrocarbons in environmental microorganisms remain elusive. However, *Tenebrio molitor* larvae (or mealworms) from a source in Beijing showed PS-degrading capacity, and a gut-PS-degrading *Exiguobacterium* spp. strain YT2 was isolated. The ubiquity of gut-microbiota-dependent PS degradation by mealworms was demonstrated later (Yang et al., 2018a, 2018b). Mealworms can also biodegrade PE (Brandon et al.,

SCIENTIFIC REPORTS 

OPEN

KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analysesReceived: 22 September 2017
Accepted: 16 March 2018
Published online: 04 April 2018Jungeun Kim¹, Jessica A. Weber², Sungwoong Jho¹, Jinho Jang^{3,4}, JeHoon Jun^{1,5}, Yun Sung Cho⁵, Hak-Min Kim^{3,4}, Hyunho Kim⁵, Yumi Kim⁵, OkSung Chung^{1,5}, Chang Geun Kim⁶, HyeJin Lee¹, Byung Chul Kim⁷, Kyudong Han⁸, InSong Koh⁹, Kyun Shik Chae⁶, Semin Lee^{3,4}, Jeremy S. Edwards¹⁰ & Jong Bhak^{1,3,4,5}

High-coverage whole-genome sequencing data of a single ethnicity can provide a useful catalogue of population-specific genetic variations, and provides a critical resource that can be used to more accurately identify pathogenic genetic variants. We report a comprehensive analysis of the Korean population, and present the Korean National Standard Reference Variome (KoVariome). As a part of the Korean Personal Genome Project (KPGP), we constructed the KoVariome database using 5.5 terabases of whole genome sequence data from 50 healthy Korean individuals in order to characterize the benign ethnicity-relevant genetic variation present in the Korean population. In total, KoVariome includes 12.7M single-nucleotide variants (SNVs), 1.7M short insertions and deletions (indels), 4K structural variations (SVs), and 3.6K copy number variations (CNVs). Among them, 2.4M (19%) SNVs and 0.4M (24%) indels were identified as novel. We also discovered selective enrichment of 3.8M SNVs and 0.5M indels in Korean individuals, which were used to filter out 1,271 coding-SNVs not originally removed from the 1,000 Genomes Project when prioritizing disease-causing variants. KoVariome health records were used to identify novel disease-causing variants in the Korean population, demonstrating the value of high-quality ethnic variation databases for the accurate interpretation of individual genomes and the precise characterization of genetic variations.

The human reference genome¹ was a milestone of scientific achievement and provides the foundation for biomedical research and personalized healthcare². The completion of the human genome marked the beginning of our concerted efforts to understand and catalogue genetic variation across human populations. The International HapMap project resolved human haplotypes into more than one million common single nucleotide polymorphisms (SNPs) in an effort to catalogue genetic variations associated with diseases³. Subsequently, other

¹Personal Genomics Institute, Genome Research Foundation, Cheongju, 28190, Republic of Korea. ²Department of Biology, University of New Mexico, Albuquerque, NM, 87131, USA. ³Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 44919, Republic of Korea. ⁴The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 44919, Republic of Korea. ⁵Geromics, Ulsan, 44919, Republic of Korea. ⁶National Standard Reference Center, Korea Research Institute of Standards and Science, Daejeon, 34113, Republic of Korea. ⁷Clinomics, Ulsan, 44919, Republic of Korea. ⁸Department of Nanobiomedical Science & BK21 PLUS NBM Global Research Center for Regenerative Medicine, Dankook University, Cheonan, 31116, Republic of Korea. ⁹Department of Physiology, College of Medicine, Hanyang University, Seoul, 04763, Republic of Korea. ¹⁰Chemistry and Chemical Biology, UNM Comprehensive Cancer Center, University of New Mexico, Albuquerque, NM, 87131, USA. Jungeun Kim, Jessica A. Weber and Sungwoong Jho contributed equally to this work. Correspondence and requests for materials should be addressed to J.S.E. (email: jsedwards@salud.unm.edu) or J.B. (email: jongbhak@genomics.org)

RESEARCH ARTICLE

Myotis rufoniger genome sequence and analyses: *M. rufoniger*'s genomic feature and the decreasing effective population size of *Myotis* bats

Youngjune Bhak^{1,2}, Yeonsu Jeon^{1,2}, Sungwon Jeon^{1,2}, Oksung Chung^{3,4}, Sungwoong Jho³, JeHoon Jun^{3,4}, Hak-Min Kim^{1,2}, Yongsoo Cho^{1,2}, Changhan Yoon^{1,5}, Seungwoo Lee⁶, Jung-Hoon Kang⁷, Jong-Deock Lim⁷, Junghwa An⁸, Yun Sung Cho^{1,2,3,*}, Doug-Young Ryu^{6,*}, Jong Bhak^{1,2,3,4,*}



OPEN ACCESS

Citation: Bhak Y, Jeon Y, Jeon S, Chung O, Jho S, Jun J, et al. (2017) *Myotis rufoniger* genome sequence and analyses: *M. rufoniger*'s genomic feature and the decreasing effective population size of *Myotis* bats. PLoS ONE 12(7): e0180418. <https://doi.org/10.1371/journal.pone.0180418>

Editor: Chongle Pan, Oak Ridge National Laboratory, UNITED STATES

Received: February 20, 2017

Accepted: May 23, 2017

Published: July 5, 2017

Copyright: © 2017 Bhak et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All sequencing files are available from the National Center for Biotechnology Information (NCBI) database (566bp: SRX2755014, 574bp: SRX2755088).

Funding: This work was supported by 'the bioinformatics marker discovery analysis system using genomic big data' Research Fund (1.150014.01) of Ulsan National Institute of Science & Technology (UNIST). It was also supported by 'Software Convergence Technology Development Program' through the Ministry of

1 The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, 2 Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, 3 Personal Genomics Institute, Genome Research Foundation, Cheongju, Republic of Korea, 4 Geromics, Ulsan, Republic of Korea, 5 Department of Biomedical Science, School of Nano-Bioscience & chemical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, 6 BK21 PLUS Program for Creative Veterinary Science Research, Research Institute for Veterinary Science, and College of Veterinary Medicine, Seoul National University, Seoul, Republic of Korea, 7 National Research Institute of Cultural Heritage, Cultural Heritage Administration, Daejeon, Republic of Korea, 8 Animal Resources Division, National Institute of Biological Resources, Incheon, Republic of Korea

* joys0406@gmail.com (YSC); dryru@snu.ac.kr (DYR); jongbhak@genomics.org (JB)

Abstract

Myotis rufoniger is a vesper bat in the genus *Myotis*. Here we report the whole genome sequence and analyses of the *M. rufoniger*. We generated 124 Gb of short-read DNA sequences with an estimated genome size of 1.88 Gb at a sequencing depth of 66× fold. The sequences were aligned to *M. brandtii* bat reference genome at a mapping rate of 96.50% covering 95.71% coding sequence region at 10× coverage. The divergence time of *Myotis* bat family is estimated to be 11.5 million years, and the divergence time between *M. rufoniger* and its closest species *M. davidii* is estimated to be 10.4 million years. We found 1,239 function-altering *M. rufoniger* specific amino acid sequences from 929 genes compared to other *Myotis* bat and mammalian genomes. The functional enrichment test of the 929 genes detected amino acid changes in melanin associated *DCT*, *SLC45A2*, *TYRP1*, and *OCA2* genes possibly responsible for the *M. rufoniger*'s red fur color and a general coloration in *Myotis*. *N6AMT1* gene, associated with arsenic resistance, showed a high degree of function alteration in *M. rufoniger*. We further confirmed that the *M. rufoniger* also has bat-specific sequences within *FSHB*, *GHR*, *IGF1R*, *TP53*, *MDM2*, *SLC45A2*, *RGS7BP*, *RHO*, *OPN1SW*, and *CNGB3* genes that have already been published to be related to bat's reproduction, lifespan, flight, low vision, and echolocation. Additionally, our demographic history analysis found that the effective population size of *Myotis* clade has been consistently decreasing since ~30k years ago. *M. rufoniger*'s effective population size was the lowest in *Myotis* bats, confirming its relatively low genetic diversity.

EVOLUTIONARY GENETICS

Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago

Veronika Siska,^{1*} Eppie Ruth Jones,^{1,2} Sungwon Jeon,³ Youngjune Bhak,³ Hak-Min Kim,³ Yun Sung Cho,³ Hyunho Kim,⁴ Kyusang Lee,⁵ Elizaveta Veselovskaya,⁶ Tatiana Balueva,⁶ Marcos Gallego-Llorente,¹ Michael Hofreiter,⁷ Daniel G. Bradley,² Anders Eriksson,¹ Ron Pinhasi,^{8,*†} Jong Bhak,^{3,4,*†‡} Andrea Manica^{1,*†}

2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. Distributed under a Creative Commons Attribution License 4.0 (CC BY).

Ancient genomes have revolutionized our understanding of Holocene prehistory and, particularly, the Neolithic transition in western Eurasia. In contrast, East Asia has so far received little attention, despite representing a core region at which the Neolithic transition took place independently ~3 millennia after its onset in the Near East. We report genome-wide data from two hunter-gatherers from Devil's Gate, an early Neolithic cave site (dated to ~7.7 thousand years ago) located in East Asia, on the border between Russia and Korea. Both of these individuals are genetically most similar to geographically close modern populations from the Amur Basin, all speaking Tungusic languages, and, in particular, to the Ulchi. The similarity to nearby modern populations and the low levels of additional genetic material in the Ulchi imply a high level of genetic continuity in this region during the Holocene, a pattern that markedly contrasts with that reported for Europe.

INTRODUCTION

Ancient genomes from western Asia have revealed a degree of genetic continuity between preagricultural hunter-gatherers and early farmers 12 to 8 thousand years ago (ka) (1, 2). In contrast, studies on southeast and central Europe indicate a major population replacement of Mesolithic hunter-gatherers by Neolithic farmers of a Near Eastern origin during the period 8.5 to 7 ka. This is then followed by a progressive "resurgence" of local hunter-gatherer lineages in some regions during the Middle/Late Neolithic and Eneolithic periods and a major contribution from the Asian Steppe later, ~5.5 ka, coinciding with the advent of the Bronze Age (3–5). Compared to western Eurasia, for which hundreds of partial ancient genomes have already been sequenced, East Asia has been largely neglected by ancient DNA studies to date, with the exception of the Siberian Arctic belt, which has received attention in the context of the colonization of the Americas (6, 7). However, East Asia represents an extremely interesting region as the shift to reliance on agriculture appears to have taken a different course from that in western Eurasia. In the latter region, pottery, farming, and animal husbandry were closely associated. In contrast, Early Neolithic societies in the Russian Far East, Japan, and Korea started to manufacture and use pottery and basketry 10.5 to 15 ka, but domesticated crops and livestock arrived several millennia later (8, 9). Because of the current lack of ancient genomes from East Asia, we do not know the extent to which this gradual Neolithic transition, which happened independently from the one taking place in western Eurasia, reflected actual

migrations, as found in Europe, or the cultural diffusion associated with population continuity.

RESULTS

Samples, sequencing, and authenticity

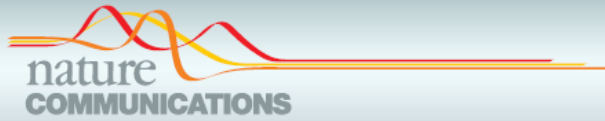
To fill this gap in our knowledge about the Neolithic in East Asia, we sequenced to low coverage the genomes of five early Neolithic burials (DevilsGate1, 0.059-fold coverage; DevilsGate2, 0.023-fold coverage; and DevilsGate3, DevilsGate4, and DevilsGate5, <0.001-fold coverage) from a single occupational phase at Devil's Gate (Chertovy Vorota) Cave in the Primorye Region, Russian Far East, close to the border with China and North Korea (see the Supplementary Materials). This site dates back to 9.4 to 7.2 ka, with the human remains dating to ~7.7 ka, and it includes some of the world's earliest evidence of ancient textiles (10). The people inhabiting Devil's Gate were hunter-fisher-gatherers with no evidence of farming; the fibers of wild plants were the main raw material for textile production (10). We focus our analysis on the two samples with the highest sequencing coverage, DevilsGate1 and DevilsGate2, both of which were female. The mitochondrial genome of the individual with higher coverage (DevilsGate1) could be assigned to haplogroup D4; this haplogroup is found in present-day populations in East Asia (11) and has also been found in Jomon skeletons in northern Japan (2). For the other individual (DevilsGate2), only membership to the M branch (to which D4 belongs) could be established. Contamination, estimated from the number of discordant calls in the mitochondrial DNA (mtDNA) sequence, was low [0.87% [95% confidence interval (CI), 0.28 to 2.37%] and 0.59% (95% CI, 0.03 to 3.753%) on nonconsensus bases at haplogroup-defining positions for DevilsGate1 and DevilsGate2, respectively. Using schmutzi (12) on the higher-coverage genome, DevilsGate1 also gives low contamination levels [1% (95% CI, 0 to 2%); see the Supplementary Materials]. As a further check against the possible confounding effect of contamination, we made sure that our most important analyses [outgroup f_3 scores and principal components analysis (PCA)] were qualitatively replicated using only reads showing evidence of postmortem damage (PMD score of at least 3) (13), although these latter results had a high level of noise due to the low coverage (0.005X for DevilsGate1 and 0.001X for DevilsGate2).

¹Department of Zoology, University of Cambridge, Downing Street, Cambridge CB23EJ, U.K. ²Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland. ³The Genomics Institute, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea. ⁴Geromics, Ulsan 44919, Republic of Korea. ⁵Clinomics Inc., Ulsan 4919, Republic of Korea. ⁶Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow, Russia. ⁷Institute for Biochemistry and Biology, Faculty for Mathematics and Natural Sciences, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany. ⁸School of Archaeology and Earth Institute, University College Dublin, Dublin, Ireland.

*Corresponding author. Email: vs389@cam.ac.uk (V.S.); ron.pinhasi@ucd.ie (R.P.); jongbhak@genomics.org (J.B.); am315@cam.ac.uk (A.M.)

†These authors contributed equally to this work.

‡Adjunct professor at Seoul National University, Seoul, Republic of Korea.



ARTICLE

Received 24 Mar 2016 | Accepted 18 Oct 2016 | Published 24 Nov 2016

DOI: 10.1038/ncomms13637

OPEN

An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes

Yun Sung Cho^{1,2,3,*}, Hyunho Kim^{4*}, Hak-Min Kim^{1,2}, Sungwoong Jho³, JeHoon Jun^{3,4}, Yong Joo Lee⁴, Kyun Shik Chae⁵, Chang Geun Kim⁵, Sangsoo Kim⁶, Anders Eriksson⁷, Jeremy S. Edwards⁸, Semin Lee^{1,2}, Byung Chul Kim^{1,2}, Andrea Manica⁷, Tae-Kwang Oh⁹, George M. Church^{10,**} & Jong Bhak^{1,2,3,4,**}

Human genomes are routinely compared against a universal reference. However, this strategy could miss population-specific and personal genomic variations, which may be detected more efficiently using an ethnically relevant or personal reference. Here we report a hybrid assembly of a Korean reference genome (KOREF) for constructing personal and ethnic references by combining sequencing and mapping methods. We also build its consensus variome reference, providing information on millions of variants from 40 additional ethnically homogeneous genomes from the Korean Personal Genome Project. We find that the ethnically relevant consensus reference can be beneficial for efficient variant detection. Systematic comparison of human assemblies shows the importance of assembly quality, suggesting the necessity of new technologies to comprehensively map ethnic and personal genomic structure variations. In the era of large-scale population genome projects, the leveraging of ethnicity-specific genome assemblies as well as the human reference genome will accelerate mapping all human genome diversity.

¹The Genomics Institute (TGI), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea. ²Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea. ³Personal Genomics Institute, Genome Research Foundation, Cheongju 28160, Korea. ⁴Geromics Inc., Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea. ⁵National Standard Reference Center, Korea Research Institute of Standards and Science, Daejeon 34113, Korea. ⁶School of Systems Biomedical Science, Soongsil University, Seoul 06978, Korea. ⁷Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. ⁸Chemistry and Chemical Biology, UNM Comprehensive Cancer Center, University of New Mexico, Albuquerque, New Mexico 87131, USA. ⁹Infection and Immunity Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Korea. ¹⁰Department of Genetics, New Research Building (NRB), Harvard Medical School, 77 Avenue Louis Pasteur, Room 238, Boston, Massachusetts 02115, USA. * These authors contributed equally to this work. ** These authors jointly supervised this work. Correspondence and requests for materials should be addressed to G.M.C. (email: gc@harvard.edu) or to J.B. (email: jongbhak@genomics.org).

Chung et al. *Genome Biology* (2015) 16:215
DOI 10.1186/s13059-015-0780-4



RESEARCH

Open Access



The first whole genome and transcriptome of the cinereous vulture reveals adaptation in the gastric and immune defense systems and possible convergent evolution between the Old and New World vultures

Oksung Chung^{1†}, Seondeok Jin^{2†}, Yun Sung Cho^{1,3†}, Jeongheui Lim⁴, Hyunho Kim³, Sungwoong Jho¹, Hak-Min Kim³, JeHoon Jun¹, HyeJin Lee¹, Alvin Chon³, Junsu Ko⁵, Jeremy Edwards⁶, Jessica A. Weber⁷, Kyudong Han^{8,9}, Stephen J. O'Brien^{10,11,12}, Andrea Manica¹³, Jong Bhak^{1,3,14*} and Woon Kee Paek^{4*}

Abstract

Background: The cinereous vulture, *Aegypius monachus*, is the largest bird of prey and plays a key role in the ecosystem by removing carcasses, thus preventing the spread of diseases. Its feeding habits force it to cope with constant exposure to pathogens, making this species an interesting target for discovering functionally selected genetic variants. Furthermore, the presence of two independently evolved vulture groups, Old World and New World vultures, provides a natural experiment in which to investigate convergent evolution due to obligate scavenging.

Results: We sequenced the genome of a cinereous vulture, and mapped it to the bald eagle reference genome, a close relative with a divergence time of 18 million years. By comparing the cinereous vulture to other avian genomes, we find positively selected genetic variations in this species associated with respiration, likely linked to their ability of immune defense responses and gastric acid secretion, consistent with their ability to digest carcasses. Comparisons between the Old World and New World vulture groups suggest convergent gene evolution. We assemble the cinereous vulture blood transcriptome from a second individual, and annotate genes. Finally, we infer the demographic history of the cinereous vulture which shows marked fluctuations in effective population size during the late Pleistocene.

Conclusions: We present the first genome and transcriptome analyses of the cinereous vulture compared to other avian genomes and transcriptomes, revealing genetic signatures of dietary and environmental adaptations accompanied by possible convergent evolution between the Old World and New World vultures.

Keywords: Cinereous vulture, Old world vulture, New world vulture, Transcriptome, Genome, Next-generation sequencing

* Correspondence: jongbhak@genomics.org; paekwk@naver.com

†Equal contributors

¹Personal Genomics Institute, Genome Research Foundation, Osong 361-951, Republic of Korea

⁴National Science Museum, Daejeon 305-705, Republic of Korea

Full list of author information is available at the end of the article



© 2015 Chung et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

