

Сборка генома обыкновенной ласки (*Mustela nivalis*) на уровне С-скаффолдов хромосомной длины

Тотиков А.^{1, 2*}, Томаровский А.^{1, 2}, Перельман П.¹, Бульонкова Т.³, Панов В.⁴, Консорциум DNA Zoo⁵, Графодатский А.¹, Кливер С.⁶

¹ Институт молекулярной и клеточной биологии СО РАН, Новосибирск, Россия

² Новосибирский государственный университет, Новосибирск, Россия

³ Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, Россия

⁴ Институт систематики и экологии животных СО РАН, Новосибирск, Россия

⁵ Центр архитектуры генома, Отделение молекулярной генетики и генетики человека, Медицинский колледж Бэйлора, Хьюстон, США

⁶ Независимый исследователь, Копенгаген, Дания

* a.totickov1@gmail.com

Ключевые слова: сборка генома; Hi-C; обыкновенная ласка; *Mustela nivalis*

Мотивация и цель: Обыкновенная ласка является широко распространенным мелким хищником, охватывающим весь Голарктический регион. Оценка генетического разнообразия различных популяций ласки вызывает большой интерес, поскольку отмечаются сокращения численности в некоторых локальных популяциях. Для более точного понимания генетического разнообразия необходимо наличие высококачественной сборки генома хромосомного уровня. Такие сборки возможно получить, используя данные о конформации хроматина (Hi-C). Используя данные Hi-C секвенирования, мы улучшили общедоступную сборку генома обыкновенной ласки на уровне отдельных скаффолдов до хромосомного уровня, а также сопоставили номенклатуру хромосомных скаффолдов с ранее полученными данными Zoo-FISH и G-бэндинга [1, 2].

Методы и алгоритмы: Общедоступная сборка генома *Mustela nivalis* (GCA_019141155.1) на уровне отдельных скаффолдов использовалась в качестве черного варианта сборки [3]. Hi-C библиотеки были получены из 6 пассажа первичной линии клеток фибробластов ушного хряща особи, отловленной в Новосибирской области. Секвенирование Hi-C библиотек проводилось на базе консорциума “DNA Zoo”. Выравнивание Hi-C данных выполнено с использованием программы Juicer v1.6 [4]. Определение гаплотипических дупликаций проводилось программой purge_dups v1.2.6 [5]. Из данных были удалены те кандидатные регионы, в которых более 90 % последовательности было вовлечено в гаплотипическую дупликацию. Сборка генома осуществлялась с помощью программы 3d-dna v180419 [6]. Ручная коррекция проводилась с использованием Juicebox v2.16.00 [7]. Целостность сборки генома оценивалась с помощью программы BUSCO v5.4.2 с использованием базы данных OrthoDB v10 Mammalia [8, 9]. Полногеномные выравнивания выполнены программой LAST v1519 [10]. Визуализация данных выполнена с использованием программного пакета MAVR (<https://github.com/mahajrod/MAVR>).

Для анализа синтенных блоков использовались сборки геномов хромосомного уровня *Martes foina* (DNA Zoo), *N. vison* (GCF_020171115.1), *M. erminea* (GCF_009829155.1), *M. nivalis*, *M. lutreola* (GCF_030435805.1), *M. putorius furo*

(DNA Zoo) и *M. nigripes* (DNA Zoo). Множественное полногеномное выравнивание было выполнено с помощью программы Progressive Cactus v2.8.0 [11]. Извлечение синтенных блоков выполнено с помощью halSynteny v2.2 [12], с указанием параметров “--minBlockSize 50000 --maxAnchorDistance 50000”. Полученные результаты были визуализированы с помощью программного пакета MACE (<https://github.com/mahajrod/MACE>).

Результаты: Сборка генома *M. nivalis* хромосомного уровня имеет размер 2.4 млрд п. н. и значение N50 138.3 млн п. н. Анализ целостности сборки генома с использованием базы данных Mammalia (всего 9226 ортологов) выявил 96 % цельных однокопийных ортологов и 3 % отсутствующих. Улучшенная сборка генома включает 21 хромосомный скаффолд (С-скаффолды), что соответствует гаплоидному набору хромосом в кариотипе *M. nivalis* ($2n=42$) [1, 2]. Наименование С-скаффолдов в сборке *M. nivalis* проведено с использованием ранее предложенной номенклатуры хромосом [1, 2]. Для этого были сопоставлены: полногеномные выравнивания сборки *M. nivalis* на сборку *Neogale vison* (GCF_020171115.1, $2n=30$), сборку *M. lutreola* (GCF_030435805.1, $2n=38$) и сборку *M. erminea* (GCF_009829155.1, $2n=44$, рис. 1B), а также результаты Zoo-FISH с использованием зондов на основе хромосом *N. vison* на хромосомы *M. lutreola* и *M. nivalis* [1], и G-banded кариотипы этих видов [2]. После идентификации С-скаффолды были отнесены к соответствующим номерам хромосом с использованием префикса “chr” (рис. 1A).

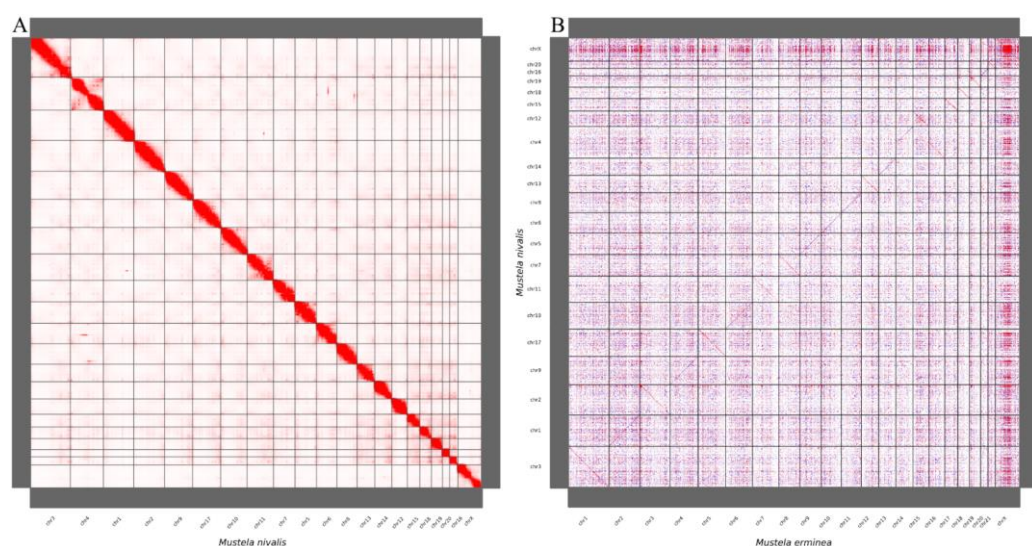


Рис. 1. (А) Карта плотности Hi-C контактов сборки генома обыкновенной ласки с 21 хромосомным скаффолдом. (В) Полногеномное выравнивание хромосомной сборки обыкновенной ласки на хромосомную сборку генома горностая (*M. erminea*)

При анализе синтенных блоков (рис. 2) обнаружены ранее цитогенетически выявленные транслокации [1]. Например, Робертсоновская транслокация двух акроцентрических хромосом, наблюдаемая у большинства видов рода *Mustela*, слияние которых образует хромосому 1 у *M. lutreola* и *M. nigripes*. Также выявлено несколько крупных инверсий размером более 1 млн п. н., которые ранее подтверждены не были.

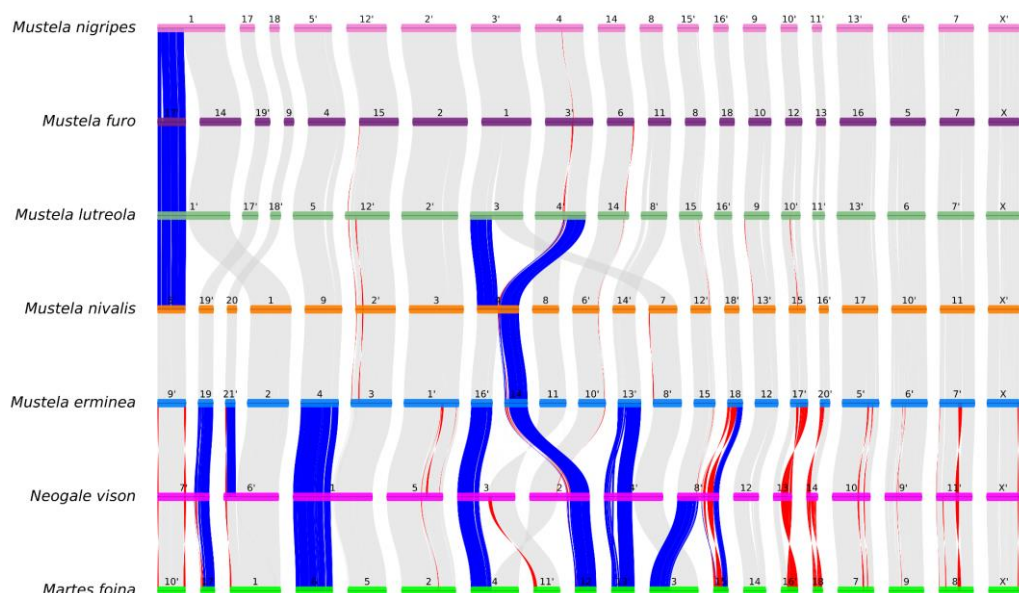


Рис. 2. Синтенные блоки между видами подсемейства Mustelinae и *Martes foina*. Инверсии и транслокации размером не менее 1 млн п.н. выделены красным и синим цветами соответственно

Выводы: Используя Hi-C данные, мы улучшили сборку генома *M. nivalis* до хромосомного уровня. Были выявлены многочисленные инверсии размером более 1 млн п.н., которые ранее определены не были. Полученная сборка генома может использоваться в различных популяционно-генетических исследованиях, позволяя получать более точные результаты.

Финансирование: За возможность проведения данной работы мы благодарим: Olga Dudchenko, Ruqayya Khan и Erez Aiden (Консорциум “DNA Zoo”, США); Karol Zub (MRI, Польша); Inês Miranda, Liliana Farelo и Jose Melo-Ferreira (CIBIO, Португалия); Юлию Бутакову (ИМКБ СО РАН, Россия). Исследование поддержано грантом РФФИ (№ 22-24-01076).

Chromosome-level genome assembly of the Least Weasel (*Mustela nivalis*)

Totikov A.^{1, 2*}, Tomarovsky A.^{1, 2}, Perelman P.¹, Bulyonkova T.³, Panov V.⁴, DNA Zoo Consortium⁵, Graphodatsky A.¹, Kliver S.⁶

¹ Institute of Molecular and Cellular Biology, SB RAS, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia

³ A.P. Ershov Institute of Informatics Systems, SB RAS, Novosibirsk, Russia

⁴ Institute of Systematics and Ecology of Animals, SB RAS, Novosibirsk, Russia

⁵ The Center for Genome Architecture, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, USA

⁶ Independent scientist, Copenhagen, Denmark

* a.totikov1@gmail.com

Key words: genome assembly; Hi-C; least weasel; *Mustela nivalis*

Motivation and Aim: The Least Weasel is a widely distributed small predator encompassing the entire Holarctic region. The assessment of genetic diversity among different populations of Least Weasels is of great interest due to observed declines in some local populations. To achieve a more accurate understanding of genetic diversity, high-quality genome assemblies at the chromosome level are necessary. Such assemblies can be obtained using chromatin conformation data (Hi-C). Utilizing Hi-C sequencing data, we have improved the publicly available genome assembly of the *M. nivalis* to the chromosome level and have also aligned the nomenclature of chromosomal scaffolds with previously obtained Zoo-FISH and G-banding data [1, 2].

Methods and Algorithms: The publicly available genome assembly of *M. nivalis* (GCA_019141155.1) at the scaffold level was utilized as a draft assembly [3]. Hi-C libraries were generated from the 6th passage of the primary fibroblast cell line derived from ear cartilage of an male individual captured in the Novosibirsk region. Hi-C library sequencing was conducted by the “DNA Zoo” consortium. Alignment of Hi-C data was performed using Juicer v1.6 software [4]. Haplotypic duplications were identified using purge_dups v1.2.6 software [5]. Candidate regions where more than 90 % of the sequence was involved in haplotypic duplication were removed. Genome assembly was carried out using 3d-dna v180419 software [6]. Manual curation was conducted using Juicebox v2.16.00 software [7]. Genome assembly completeness was assessed using BUSCO v5.4.2 software with the OrthoDB v10 Mammalia database [8, 9]. Whole-genome alignments were performed using LAST v1519 software [10]. Data visualization was conducted using the MAVR software package (<https://github.com/mahajrod/MAVR>). The synteny blocks analysis utilized chromosome-level genome assemblies at the *Martes foina* (DNA Zoo), *N. vison* (GCF_020171115.1), *M. erminea* (GCF_009829155.1), *M. nivalis*, *M. lutreola* (GCF_030435805.1), *M. putorius furo* (DNA Zoo), and *M. nigripes* (DNA Zoo). Multiple whole-genome alignments were conducted using Progressive Cactus v2.8.0 [11]. Syntenic blocks were extracted using halSynteny v2.2 [12] with parameters set to “--minBlockSize 50000 --maxAnchorDistance 50000”. The resulting outcomes were visualized using the MACE software package (<https://github.com/mahajrod/MACE>).

Results: The improved genome assembly has a size of 2.4 Gbp and an N50 value of 138.3 Mbp. BUSCO analysis for the *M. nivalis* assembly, we identified 96 % complete and single-copy BUSCOs from the Mammalia database (9,226 BUSCOs), with only 3 % missing. The improved genome assembly consists of 21 chromosomal scaffolds (C-scaffolds), corresponding to the haploid chromosome set in the karyotype of *M. nivalis* ($2n=42$) [1, 2]. The nomenclature of C-scaffolds in the *M. nivalis* assembly was established using previously proposed chromosome nomenclature [1, 2]. This involved comparing whole-genome alignments of the *M. nivalis* assembly to the assembly of *Neogale vison* (GCF_020171115.1, $2n=30$), the assembly of *M. lutreola* (GCF_030435805.1, $2n=38$), and the assembly of *M. erminea* (GCF_009829155.1, $2n=44$), as well as the results of Zoo-FISH probes based on *N. vison* chromosomes on *M. lutreola* and *M. nivalis* chromosomes [1], and G-banded karyotypes of these species [2]. Following the identification of C-scaffolds, they were assigned to corresponding chromosome numbers using the prefix “chr” (Fig. 1).

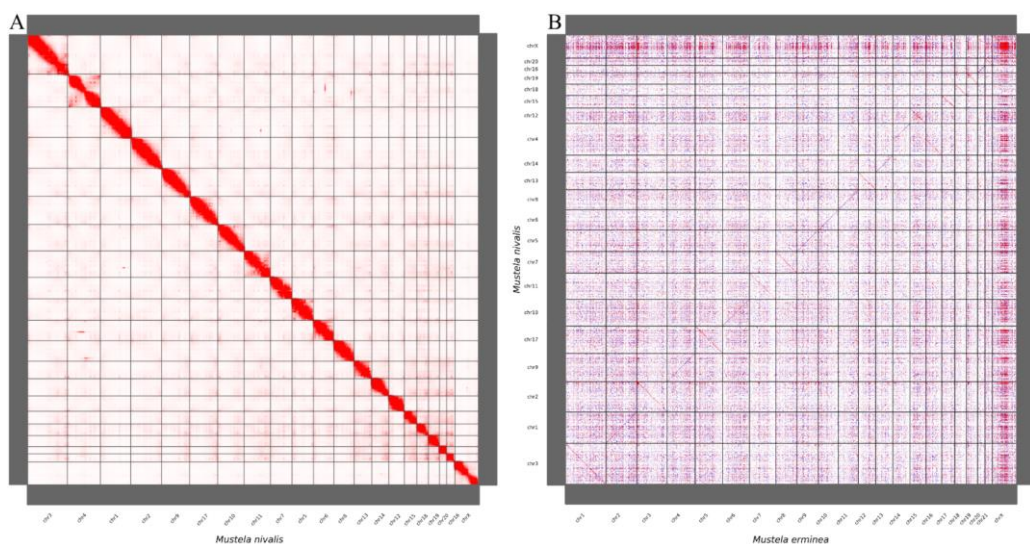


Fig. 1. (A) Hi-C contact density map of the Least Weasel genome assembly showing 21 chromosome length scaffolds. (B) Whole genome alignment of the Least Weasel chromosome length assembly to the ermine's (*M. erminea*) assembly

During the analysis of synteny blocks (Fig. 2), previously cytogenetically identified translocations were detected [1]. For instance, the Robertsonian translocation involves two acrocentric chromosomes, observed in most species of the genus *Mustela*, the fusion of which forms chromosome 1 in *M. lutreola* and *M. nigripes*. Additionally, several large inversions exceeding 1 Mbp were identified.

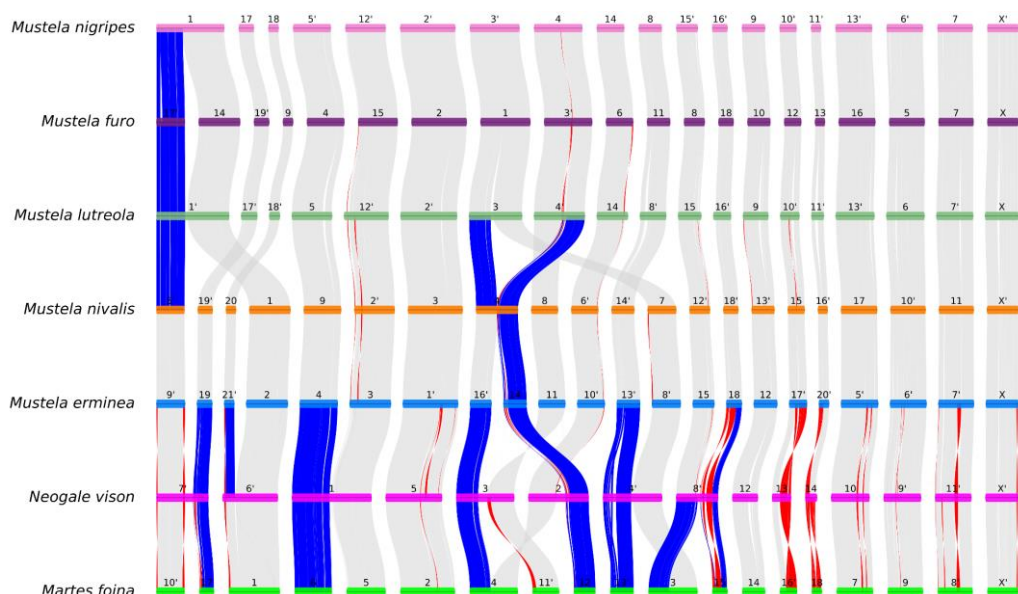


Fig. 2. Synteny blocks between species of the Mustelinae subfamily and *Martes foina*. Inversions and translocations of 1 Mbp and larger are highlighted in red and blue, respectively

Conclusion: Using Hi-C data, we have successfully improved the genome assembly of *M. nivalis* to the chromosome level. We detected numerous inversions larger than 1 Mbp that were previously undetected. The resulting genome assembly can be utilized in various population genetics studies, enabling more precise results to be obtained.

Funding: We acknowledge for making this genome assembly possible: Olga Dudchenko, Ruqayya Khan and Erez Aiden (Consortium “DNA Zoo”, USA); Karol Zub (MRI, Poland); Inês Miranda, Liliana Farelo, Jose Melo-Ferreira (CIBIO, Portugal); and Yulia Butakova (IMCB SB RAS, Russia). The study is supported by the RSF grant (No. 22-24-01076).

Список литературы/References

1. Graphodatsky A.S., Yang F., Perelman P.L. et al. Comparative molecular cytogenetic studies in the order Carnivora: Mapping chromosomal rearrangements onto the phylogenetic tree. *Cytogenet Genome Res.* 2002;96(1-4):137-145. doi 10.1159/000063032
2. Graphodatsky A., Perelman P., O'Brien S.J. Atlas of Mammalian Chromosomes. John Wiley & Sons, Incorporated, 2020. doi 10.1002/9781119418061
3. Miranda I., Giska I., Farelo L. et al. Museomics dissects the genetic basis for adaptive seasonal colouration in the least weasel. *Mol Biol Evol.* 2021;38(10):4388-4402. doi 10.1093/molbev/msab177
4. Durand N.C., Shamim M.S., Machol I. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 2016;3(1):95-98. doi 10.1016/j.cels.2016.07.002
5. Guan D., McCarthy S.A., Wood J., Howe K., Wang Y., Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* 2020;36(9):2896-2898. doi 10.1093/bioinformatics/btaa025
6. Dudchenko O., Batra S.S., Omer A.D. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* 2017;356(6333):92-95. doi 10.1126/science.aal3327
7. Durand N.C., Robinson J.T., Shamim M.S. et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* 2016;3(1):99-101. doi 10.1016/j.cels.2015.07.012
8. Manni M., Berkeley M.R., Seppey M. et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38(10):4647-4654. doi 10.1093/molbev/msab199
9. Kriventseva E.V., Kuznetsov D., Tegenfeldt F. et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 2019;47(D1):D807-D811. doi 10.1093/nar/gky1053
10. Kielbasa S.M., Wan R., Sato K., Horton P., Frith M.C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* 2011;21(3):487-493. doi 10.1101/gr.113985.110
11. Armstrong J., Hickey G., Diekhans M. et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature.* 2020;587(7833):246-251. doi 10.1038/s41586-020-2871-y
12. Krashennikova K., Diekhans M., Armstrong J. et al. halSynteny: a fast, easy-to-use conserved syntenic block construction method for multiple whole-genome alignments. *GigaScience.* 2020;9(6):giaa047. doi 10.1093/gigascience/giaa047