

## Актуальность интеграции данных цитогенетики и сборок геномов хромосомного уровня на примере различных групп позвоночных

Романенко С.<sup>1\*</sup>, Беклемишева В.<sup>1</sup>, Перельман П.<sup>1</sup>, Андреюшкова Д.<sup>1</sup>,  
Марченко С.<sup>2</sup>, Лемская Н.<sup>1</sup>, Билтуева Л.<sup>1</sup>, Кливер С.<sup>3</sup>, Трифонов В.<sup>1</sup>,  
Графодатский А.<sup>1</sup>

<sup>1</sup> Институт молекулярной и клеточной биологии СО РАН, Новосибирск, Россия

<sup>2</sup> Новосибирский государственный университет, Новосибирск, Россия

<sup>3</sup> Независимый исследователь, Копенгаген, Дания

\* [rosa@mcb.nsc.ru](mailto:rosa@mcb.nsc.ru)

**Ключевые слова:** дифференциальная окраска хромосом; гетерохроматин; кариотип; скаффолд; флуоресцентная гибридизация *in situ*; хромосомные перестройки; центромера

**Мотивация и цель:** Сборки до уровня хромосом являются золотым стандартом для секвенирования всего генома *de novo* [1–3]. Получение высококачественных эталонных геномов позвоночных, собранных до хромосомного уровня, привело к значительным результатам в функциональной, сравнительной и популяционной геномике [4, 5]. Сложными этапами при построении эталонного референсного генома являются правильная ориентация контигов и скаффолдов по отношению друг к другу и сборка суперскаффолдов в структурированные хромосомные скаффолды, а также секвенирование и позиционирование участков генома с повторяющимися последовательностями или с сегментными дупликациями [6–8]. Многие из этих проблем преодолеваются при интеграции данных секвенирования генома и данных цитогенетики. Комбинированный подход, обеспечивающий сборку генома до уровня хромосом, гораздо более успешен в ответе на фундаментальные биологические вопросы, чем одни лишь геномные или цитогенетические подходы [9–11]. Зачастую сопоставление данных секвенирования генома и цитогенетических данных не проводится, что отражается в несогласованности нумерации хромосом даже для видов, для которых имеются стандартизированные кариотипы, к неверной ориентации скаффолдов, к недооценке размеров хромосом из-за неучетов блоков гетерохроматина и т. д. Целью данной работы является привлечение внимания к важности понимания и применения комбинированного подхода при выполнении сборок геномов до хромосомного уровня.

**Методы и алгоритмы:** В работе использовались методы классической, дифференциальной и молекулярной цитогенетики [12–15]. Микродиссекция проводилась для получения хромосоמו- и районоспецифичных библиотек [16]. Приготовление библиотек для секвенирования выполнялось, как описано ранее [17]. Для анализа данных секвенирования применялись различные подходы [17–20].

**Результаты:** Путем парного выравнивания мы идентифицировали и визуализировали синтенные блоки, общие для морской свинки (*Cavia porcellus*), голого землекопа (*Heterocephalus glaber*) и человека. Высокое качество сборки генома человека позволило определить положение центромер при сравнении синтений.

Мы объединили мелкие синтении в блоки размером не менее 5 м.п.о. для сопоставления данных сравнительной цитогенетики и сборок геномов, полученных методом анализа данных секвенирования библиотек Hi-C, что позволило привязать геномные скаффолды к отдельным хромосомам и ориентировать их относительно центромер и блоков гетерохроматина. Данные Hi-C позволили закрыть все пробелы на сравнительных картах и выявить дополнительные перестройки, отличающие кариотипы трех видов. В результате мы интегрировали биоинформатические и цитогенетические данные и скорректировали предыдущие сравнительные карты и сборки генома морской свинки, голого землекопа и человека [19].

В кариотипе стерляди (*Acipenser ruthenus*, ARUT) из 120 хромосом цитогенетически легко идентифицируются первые 14–17 пар и пары мелких хромосом, несущие гены ЯОР [21]. Методом микродиссекции мы получили библиотеки отдельных хромосом стерляди. Эти библиотеки были помечены и использованы в качестве зондов для FISH, а также секвенированы и выровнены на сборку генома стерляди хромосомного уровня (GCF\_902713425.1/), чтобы увязать цитогенетические и биоинформатические данные. Анализ результатов FISH показал, что геном стерляди представляет сложную мозаичную структуру, содержащую диплоидные и тетраплоидные сегменты [22]. При сопоставлении молекулярно-цитогенетических и биоинформатических данных было получено хорошее или приемлемое соответствие результатов для первых двадцати пар хромосом. Для четырех пар хромосом было выявлено существенное несоответствие между физическими размерами хромосом (ARUT17, 21, 22 и 25) и размерами соответствующих им скаффолдов (24, 15, 29 и 18), объясняемое, скорее всего, неполной сборкой. Для большого по размеру скаффолда 11 не удалось установить соответствия с какой-либо хромосомной парой. По-видимому, это артефактный результат.

При сравнении геномов красноухой пресноводной черепахи (*Trachemys scripta elegans*, TSE) и логгерхеда (*Caretta caretta*, CCAR) инструментом D-GENIES были выявлены межхромосомные перестройки, затрагивающие хромосомы TSE2=CCAR2/23/2, TSE4=CCAR6/20 и TSE16=CCAR26/27, также была показана обратная ориентация скаффолдов для гомологичных хромосом TSE1 и CCAR1, TSE4 и CCAR6, TSE9 и CCAR9 и др. Молекулярно-цитогенетическая локализация наборов хромосомоспецифичных проб TSE и коробчатой черепахи (*Terrapene carolina*, TCA) на хромосомах TSE и CCAR хотя и подтвердила наличие трех межхромосомных перестроек, различающих кариотипы TSE и CCAR, но указала на то, что сборка хромосомного уровня не привязана к цитогенетической карте: так, CCAR6 и TSE16 в сборках соответствуют физически CCAR5 и TSE12. Использование микродиссекционных районспецифичных проб для сравнительного хромосомного пэинтинга выявило идентичный порядок расположения гомологичных сегментов, например, в гомологичных хромосомах TSE1 и CCAR1, указав на неверную ориентацию скаффолдов в хромосомных сборках.

Сравнительный хромосомный пэинтинг с пробами домашней собаки продемонстрировал одинаковый порядок гомологичных фрагментов для всех аутосом ластиногих. Высокий уровень консерватизма аутосом был подтвержден при сборке генома байкальской нерпы (*Pusa sibirica*, PSIB) методом Hi-C и сравнении его с геномами пестрой нерпы (*Phoca larga*, PLAR) и обыкновенного

тюленя (*Phoca vitulina*, PVIT) [20]. Было выявлено множество коротких инверсий, и только три инверсии размером около 1 м.п.н. на PLAR6 и PVIT6. Размер выявленных внутривнутрихромосомных перестроек ниже уровня разрешающей способности метода FISH, что подчеркивает важность проведения сравнительного биоинформатического анализа геномов. Однако полученные данные по сборке генома байкальской нерпы до уровня хромосом не содержат сведений о положении центромера, что не позволяет установить, в каком именно плече метацентрической хромосомы находятся выявленные перестройки.

**Выводы:** Интеграция цитогенетических и биоинформатических данных для корректных сборок геномов хромосомного уровня представителей разных таксонов позвоночных имеет огромное значение, позволяя корректировать ориентацию скаффолдов и устанавливать соответствие с хромосомами кариотипов, описанных молекулярно-цитогенетическими методами.

## Relevance of integration of cytogenetics data and chromosome-length genome assemblies at the example of various vertebrates' groups

Romanenko S.<sup>1\*</sup>, Beklemisheva V.<sup>1</sup>, Perelman P.<sup>1</sup>, Andreyushkova D.<sup>1</sup>, Marchenko S.<sup>2</sup>, Lemskaya N.<sup>1</sup>, Biltueva L.<sup>1</sup>, Kliver S.<sup>3</sup>, Trifonov V.<sup>1</sup>, Graphodatsky A.<sup>1</sup>

<sup>1</sup> Institute of Molecular and Cellular Biology, SB RAS, Novosibirsk, Russia

<sup>2</sup> Novosibirsk State University, Novosibirsk, Russia

<sup>3</sup> Independent researcher, Copenhagen, Denmark

\* rosa@mcb.nsc.ru

**Key words:** banding; centromere; chromosome rearrangements; fluorescent *in situ* hybridization; heterochromatin; karyotype; scaffold

**Motivation and Aim:** Chromosome-length genome assemblies are the gold standard for *de novo* whole genome sequencing [1–3]. The availability of high-quality vertebrate reference genomes led to significant advances in functional, comparative and population genomics [4, 5]. Problematic steps in constructing a reliable reference genome are the correct orientation of contigs and scaffolds relative to each other and the assembly of superscaffolds into structured chromosomal scaffolds, as well as the sequencing and positioning of genomic regions with repeated sequences or segmental duplications [6–8]. Many of the above problems are overcome by integrating genome sequencing data with cytogenetics data. A combined approach that assembles the genome down to the chromosome level is much more successful in answering fundamental biological questions than genomic or cytogenetic approaches alone [9–11]. Unfortunately, integration of genome sequencing data with cytogenetics data is often not carried out, which leads to inconsistency in chromosome numbering even for species for which standardized karyotypes are available, incorrect orientation of scaffolds, incorrect assessment of chromosome sizes due to failure to take into account heterochromatic blocks, and other problems. The purpose of this work is to draw attention to the

importance of understanding and applying a combined approach when performing chromosome-length genome assemblies.

**Methods and Algorithms:** The methods of classical, differential and molecular cytogenetics were used [12–15]. Microdissection was carried out to obtain chromosome- and region-specific libraries [16]. Preparation of sequencing libraries was performed as previously described [17]. Analysis of sequencing data was carried out using various approaches [17–20].

**Results:** By pairwise alignment, we identified and visualized syntenic blocks shared by two rodent species: the guinea pig (*Cavia porcellus*), the naked mole rat (*Heterocephalus glaber*), and humans. The high quality of the human genome assembly made it possible to determine the position of centromeres when comparing syntenies. We grouped small syntenies into blocks of at least 5 Mbp to align comparative cytogenetics data and genome assemblies obtained by analyzing sequencing data from Hi-C libraries, which made it possible to link genomic scaffolds to individual chromosomes and orient them relative to centromeres and heterochromatic blocks. Hi-C data made it possible to close all gaps in the comparative maps and identify additional rearrangements that distinguish the karyotypes of the three species. As a result, we integrated bioinformatics and cytogenetic data and adjusted previous comparative maps and genome assemblies of the guinea pig, naked mole rat, and human [19].

In the sterlet, a Eurasian sturgeon fish, (*Acipenser ruthenus*, ARUT) karyotype of 120 chromosomes, the first 14–17 pairs and pairs of small chromosomes carrying NOR genes are easily identified cytogenetically [21]. Using the microdissection, we obtained libraries of individual sterlet chromosomes. These libraries were labeled and used as probes for FISH, and sequenced and aligned to the chromosomal level sterlet genome assembly (GCF\_902713425.1/) to link cytogenetic and bioinformatics data. Analysis of FISH results showed that the sterlet genome represents a complex mosaic structure containing diploid and tetraploid segments [22]. When comparing molecular cytogenetic and bioinformatics data, good or acceptable agreement of results was obtained for chromosomes ARUT1–10, 13, 14, as well as the next twenty pairs of chromosomes. For four pairs of chromosomes, a significant discrepancy was detected between the physical sizes of the chromosomes (ARUT17, 21, 22 and 25) and the sizes of their corresponding scaffolds (24, 15, 29 and 18), most likely explained by incomplete assembly. For the large scaffold 11, it was not possible to establish a correspondence with any chromosomal pair.

When comparing the genomes of two turtles: the red-eared slider (*Trachemys scripta elegans*, TSE) and the loggerhead sea turtle (*Caretta caretta*, CCAR) using the D-GENIES tool, interchromosomal rearrangements TSE2=CCAR2/23/2, TSE4=CCAR6/20, and TSE16=CCAR26/27 were identified, the reverse orientation of scaffolds was also shown for homologous chromosomes TSE1 and CCAR1, TSE4 and CCAR6, TSE9 and CCAR9, etc. Molecular cytogenetic localization of sets of chromosome-specific probes TSE and the eastern box turtle (*Terrapene carolina*, TCA) on the chromosomes of TSE and CCAR, although confirmed the presence of the three interchromosomal rearrangements distinguishing the TSE and CCAR karyotypes, but indicated that the chromosome-length genome assembly is not tied to the cytogenetic map: thus, CCAR6 and TSE16 in the assembly physically correspond to CCAR5 and TSE12. The use of microdissection region-specific probes for comparative chromosomal painting revealed an identical order of arrangement of homologous segments, for

example, in the homologous chromosomes of TSE1 and CCAR1, indicating the incorrect orientation of scaffolds in the turtles' chromosomal assemblies.

Comparative chromosomal painting with the domestic dog probes demonstrated the same order of homologous fragments for all pinniped autosomes (Carnivora, Mammalia). A high level of autosomal conservatism was confirmed by assembling the genome of the Baikal seal (*Pusa sibirica*, PSIB) using the Hi-C method and comparing it with the genomes of the pied seal (*Phoca larga*, PLAR) and the common seal (*Phoca vitulina*, PVIT) [20]. Many short inversions and three inversions approximately 1 Mb in size on PLAR6 and PVIT6 were identified. The size of the identified intrachromosomal rearrangements is below the resolution level of FISH, which emphasizes the importance of conducting comparative bioinformatics analysis of genomes. However, the data obtained on the assembly of the Baikal seal genome to the chromosome level do not contain information about the position of the centromeres, which does not allow us to establish in which arm of the metacentric chromosome the identified rearrangements were located.

**Conclusion:** Integration of cytogenetic and bioinformatics data for correct chromosome-length genome assemblies is of great importance, making it possible to correct the orientation of scaffolds and to match chromosome scaffolds with chromosomes described by molecular cytogenetic methods in representatives of different vertebrate taxa.

### Список литературы/References

1. Rhie A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021;592(7856):737-746. doi 10.1038/s41586-021-03451-0
2. Kim J. et al. Reference-assisted chromosome assembly. *Proc Natl Acad Sci USA*. 2013;110(5):1785-1790. doi 10.1073/pnas.1220349110
3. A reference standard for genome biology. *Nat Biotechnol*. 2018;36(12):1121. doi 10.1038/nbt.4318
4. Garner B.A. et al. Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application. *Trends Ecol Evol*. 2016;31(2):81-83. doi 10.1016/j.tree.2015.10.009
5. Supple M.A., Shapiro B. Conservation of biodiversity in the genomics era. *Genome Biol*. 2018;19:13. doi 10.1186/s13059-018-1520-3
6. Zimin A.V. et al. Mis-assembled "segmental duplications" in two versions of the *Bos taurus* genome. *PLoS One*. 2012;7(8):e42680. doi 10.1371/journal.pone.0042680
7. Hills M. et al. Construction of whole genomes from scaffolds using single cell strand-seq data. *Int J Mol Sci*. 2021;22(7):3617. doi 10.3390/ijms22073617
8. Luo J. et al. A comprehensive review of scaffolding methods in genome assembly. *Brief Bioinform*. 2021;22(5):bbab033. doi 10.1093/bib/bbab033
9. Deakin J.E. et al. Chromosomics: Bridging the Gap between Genomes and Chromosomes. *Genes (Basel)*. 2019;10(8):627. doi 10.3390/genes10080627
10. Lewin H.A. et al. Precision nomenclature for the new genomics. *GigaScience*. 2019;8(8):giz086. doi 10.1093/gigascience/giz086
11. Iannucci A. et al. Bridging the gap between vertebrate cytogenetics and genomics with Single-Chromosome Sequencing (ChromSeq). *Genes (Basel)*. 2021;12(1):124. doi 10.3390/genes12010124
12. Lemskaya N.A. et al. A combined banding method that allows the reliable identification of chromosomes as well as differentiation of AT- and GC-rich heterochromatin. *Chromosome Res*. 2018;26(4):307-315. doi 10.1007/s10577-018-9589-9
13. Seabright M. A rapid banding technique for human chromosomes. *Lancet*. 1971;2(7731):971-972. doi 10.1016/s0140-6736(71)90287-x
14. Sumner A.T. A simple technique for demonstrating centromeric heterochromatin. *Exp Cell Res*. 1972;75(1):304-306. doi 10.1016/0014-4827(72)90558-7
15. Yang F., Graphodatsky A.S. Animal probes and ZOO-FISH. In: Liehr T. (Ed.). *Fluorescence In Situ Hybridization (FISH)*. Springer Protocols Handbooks, 2017. doi 10.1007/978-3-662-52959-1\_42
16. Romanenko S., Trifonov V. Generation of microdissection-derived painting probes from single copy chromosomes. In: Liehr T. (Ed.). *Cytogenetics and Molecular Cytogenetics*. CRC Press, 2022. doi 10.1201/9781003223658

17. Andreyushkova D.A. et al. Next generation sequencing of chromosome-specific libraries sheds light on genome evolution in paleotetraploid sterlet (*Acipenser ruthenus*). *Genes (Basel)*. 2017;8(11):318. doi 10.3390/genes8110318
18. Makunin A.I. et al. Contrasting origin of B chromosomes in two cervids (Siberian roe deer and grey brocket deer) unravelled by chromosome-specific DNA sequencing. *BMC Genomics*. 2016;17(1):618. doi 10.1186/s12864-016-2933-6
19. Romanenko S.A. et al. Integration of fluorescence *in situ* hybridization and chromosome-length genome assemblies revealed synteny map for guinea pig, naked mole-rat, and human. *Sci Rep*. 2023;13(1):21055. doi 10.1038/s41598-023-46595-x
20. Yakupova A. et al. Chromosome-length assembly of the Baikal seal (*Pusa sibirica*) genome reveals a historically large population prior to isolation in Lake Baikal. *Genes*. 2023;14(3):619. doi 10.3390/genes14030619
21. Romanenko S.A. et al. Segmental paleotetraploidy revealed in sterlet (*Acipenser ruthenus*) genome by chromosome painting. *Mol Cytogenet*. 2015;8:90. doi 10.1186/s13039-015-0194-8
22. Trifonov V.A. et al. Evolutionary plasticity of acipenseriform genomes. *Chromosoma*. 2016;125:661-668. doi 10.1007/s00412-016-0609-2