

Применение технологий параллельного и нанопорового секвенирования для изучения структурной организации митохондриального генома сои (*Glycine max* L. Merr.)

Александрович В.В.*, Синявская М.Г.

Институт генетики и цитологии НАН Беларуси, Минск, Беларусь

* valeria.alexandrovich@gmail.com

Ключевые слова: митохондриальный геном; соя; полные нуклеотидные последовательности; секвенирование нового поколения; нанопоровое секвенирование

Мотивация и цель: Концепция существования митохондриального генома высших растений в виде «master circle» или мастер-хромосомы ставится под сомнение в последние годы [1]. Все больше доказательств появляется в пользу динамической структуры растительной митохондриальной ДНК (мтДНК), которая представлена в клетке множеством линейных и кольцевых молекул, возникающим в процессе рекомбинации и перестроек генома [2].

Митохондриальный геном сои (*Glycine max* L. Merr) считается одним из сложнейших геномов среди культурных растений. В геноме *G. max* присутствует необычайно высокая доля длинных повторов, рекомбинация по которым приводит к возникновению, предположительно, более тысячи разных сублимонов [3].

В базах данных нуклеотидных последовательностей имеется три сборки мтДНК культурной сои: сортов Aigianhuang [4] и Zhonghuang 13 [5], митохондриальный геном которых, согласно авторам, представляет собой единственную кольцевую молекулу, и мтДНК сорта Williams 82 [3], существующей, предположительно, в виде двух кольцевых хромосом. Однако данные последовательности мтДНК не отражают возможного состояния генома митохондрий в живой клетке.

В данном исследовании была поставлена цель изучить полную нуклеотидную последовательность митохондриального генома культурной сои (*G. max*) с использованием двух методов секвенирования с учетом ее возможной лабильной структуры *in vivo*.

Методы и алгоритмы: Для сборки митохондриального генома *G. max* использовалось сочетание параллельного (Illumina MiSeq) и нанопорового (MinION, Oxford Nanopore Technologies) секвенирования. Полученные после параллельного секвенирования FASTQ-файлы были обработаны Trimmomatic, далее были собраны контиги с использованием SPAdes, контиги визуализировались с помощью Bandage. Combinator-FQ использовался для предварительной оценки копийности контигов, основываясь на их покрытии, и порядка их расположения.

Контиги из длинных прочтений нанопорового секвенирования были получены с помощью flye и canu. Выравнивание контигов параллельного секвенирования на контиги длинных прочтений (SnapGene) было использовано для оценки возможного расположения блоков мтДНК относительно друг друга и идентификации повторов в последовательности митохондриального генома.

Результаты: В результате обработки данных NGS было получено 23 контига размером от 196 до 65372 п.н., различного покрытия и GC-состава (от 36 до 46 %). Один из контигов образует кольцевую молекулу размером 62661 п.н., не имеющую областей гомологии с остальными последовательностями, что говорит о существовании митохондриального генома сои в виде как минимум двух хромосом, не рекомбинирующих друг с другом. Замкнутость данного контига подтверждается также наличием длинных нанопоровых прочтений на границе соединения его в кольцо.

Выравнивание остальных 22 контигов на нанопоровые прочтения показало их существование в виде “блоков” мтДНК, расположенных в различной последовательности и ориентации относительно друг друга.

Контиги, имеющее высокое по сравнению с остальными покрытие, представлены в геноме митохондрий в виде двух и более копий. Предположительно, данные многокопийные участки размером от 440 до 12042 п.н. образуют рекомбинационные повторы, которые и обеспечивают множество обнаруженных конформаций мтДНК.

В мтДНК сои обнаружены инверсии размером до 183678 п.н., делеции и вставки участков генома. Один из контигов размером в 15259 п.н. отличается наличием двух открытых рамок считывания, кодирующих ДНК-полимеразу группы В и РНК-полимеразу, что имеет некоторое сходство с обнаруженными ранее в митохондриях растений линейными плазмидами [6]. В случае сои данная плазида не имеет длинных инвертированных повторов и встроена в мтДНК. Два контига (196 п.н. и 1318 п.н.), по предварительным данным, образуют блок мтДНК, существующий во встроеном в хромосому виде и одновременно с этим находится в митохондриях в виде автономной кольцевой молекулы. Природа данного участка на данный момент находится под вопросом.

Всего, если представить одну из хромосом мтДНК в виде «master circle», митохондриальный геном сои имеет общий размер в 434295 п.н., GC-состав 44,94% и организован в виде двух хромосом, что соотносится с данными Liu и соавторов (сборка мтДНК сорта Williams 82) [3].

Выводы: Митохондриальный геном сои представлен *in vivo* минимум двумя молекулами, одна из которых имеет консервативную структуру, а вторая подвержена постоянным рекомбинационным перестройкам благодаря наличию длинных повторов в последовательности мтДНК и существует в клетке в виде разнообразных субгеномных молекул. Сочетание технологий Illumina и Oxford Nanopore позволило с большей уверенностью говорить о структуре мтДНК *G. max* и показало себя как перспективный метод сборки сложных митохондриальных геномов растений.

На данный момент стоит задача создания и опубликования в базе данных NCBI GenBank нуклеотидной последовательности митохондриального генома *G. max*, отражающей копийность блоков мтДНК и особенности ее структуры. Также в дальнейших исследованиях планируется изучить выявленные плазмидо-подобные молекулы.

Финансирование: Исследование поддержано ГПНИ «Биотехнологии 2» 2021–2025 гг. подпрограммы «Геномика, эпигеномика, биоинформатика» (рег. № НИОКТР 20210443).

Application of parallel and nanopore sequencing technologies for the study of soybeans (*Glycine max* L. Merr.) mitochondrial genome structural organization

Aleksandrovich V.V.*, Siniauskaya M.G.

Institute of Genetics and Cytology, NAS, Minsk, Belarus

* valeria.alexandrovich@gmail.com

Key words: mitochondrial genome; soybean; complete nucleotide sequences; next-generation sequencing; nanopore sequencing

Motivation and Aim: The master circle model of higher plant mitochondrial genome has been questioned in recent years [1]. More and more evidence is emerging in support of dynamic structure of plant mitochondrial DNA (mtDNA), which in the living cell takes the form of a variety of linear and circular molecules that arise during the process of recombination and genome rearrangements [2].

The mitochondrial genome of soybean (*Glycine max* L. Merr) is considered to be one of the most complex genomes among cultivated plants. The *G. max* mtDNA contains an unusually high number of long repeats, the recombination of which leads to the formation of more than a thousand different subgenome molecules [3].

Nucleotide sequence databases contain three assemblies of mtDNA of cultivated soybean. According to the authors, the mitochondrial genome of cultivars Aiganhuang [4] and Zhonghuang 13 [5] is the single circular molecule, and mtDNA of cultivar Williams 82 [3] exists in the form of two circular chromosomes. However, these assemblies do not reflect the possible state of the mitochondrial genome in the living cell.

The goal of this study was to investigate the complete nucleotide sequence of cultivated soybeans mitochondrial genome using two sequencing methods. Also the possible labile structure of mtDNA *in vivo* was taken in consideration.

Methods and Algorithms: A combination of parallel (Illumina MiSeq) and nanopore (MinION, Oxford Nanopore Technologies) sequencing methods was used to assemble the *G. max* mitochondrial genome. The FASTQ files obtained after next-generation sequencing were processed by Trimmomatic, after that contigs were assembled using SPAdes. Contigs were visualized using Bandage. Combinator-FQ script was used to preevaluate the order and copy number of contigs based on their coverage.

Contigs from nanopore sequencing long reads were obtained using flye and canu. Alignment of parallel sequencing contigs to long-read contigs (SnapGene) was used to estimate the possible order of mtDNA blocks and identify repeats in the mitochondrial genome sequence.

Results: As a result of NGS data processing 23 contigs ranging in size (from 196 to 65372 bp) and GC-content (36 to 46%) were obtained. One of the contigs forms a circular molecule of 62661 bp in size, which has no homologous regions with other sequences, suggesting the existence of the soybean mitochondrial genome in the form of at least two chromosomes that do not recombine with each other. The presence of long nanopore reads that contain the junction of two ends of this contig also confirms the formation of one circular molecule.

Alignment of the remaining 22 contigs to nanopore reads showed their existence in the form of blocks of mtDNA located in different order and orientation. Contigs with higher coverage are represented in the mitochondrial genome by two or more copies. Presumably, these multicopy regions ranging in size from 440 to 12042 bp form recombination repeats, which provoke the formation of multiple mtDNA conformation variants that were detected.

Inversions up to 183678 bp in size, deletions and insertions of genomic regions were observed in soybean mtDNA. One of the contigs 15259 bp long is notable for the presence of two open reading frames encoding group B DNA polymerase and RNA polymerase, which is of some similarity with linear plasmids previously described in plant mitochondria [6]. In the case of soybean, this plasmid does not have long inverted repeats and is integrated into mtDNA. Two contigs (196 bp and 1318 bp), according to preliminary data, form a mtDNA block that was found to be integrated into the chromosome and, at the same time, to be located in mitochondria as an autonomous circular molecule. The nature of this mtDNA block is currently in question.

Overall, the soybean mitochondrial genome has a total size of 434295 bp, a GC-content of 44,94% and is organized in the form of two chromosomes, which correlates with the data of Liu et al. (assembly of mtDNA of Williams 82 cultivar) [3].

Conclusion: The soybean mitochondrial genome is represented *in vivo* by at least two molecules, one of which has a conservative structure, and the second is subjected to constant recombinational rearrangements due to the presence of long repeats in the mtDNA sequence. This chromosome exists in the cell in the form of various subgenomic molecules. The combination of Illumina and Oxford Nanopore technologies allowed us to properly investigate the structure of *G. max* mtDNA and showed itself to be a promising method for assembling complex plant mitochondrial genomes.

At present, our goal is to obtain and publish in the NCBI GenBank database the full nucleotide sequence of the cultivated soybeans mitochondrial genome, which reflects the copy number and structural peculiarities of mtDNA blocks. We are also planning to investigate the nature of detected plasmid-like molecules in further studies.

Funding: The study is supported by the State Scientific Research Program «Biotechnologies 2» 2021–2025, Subprogram «Genomics, epigenomics, bioinformatics» (No. 20210443).

Список литературы/References

1. Михайлова Ю.В., Терентьева Л.Ю. Гигантские митохондриальные геномы высших растений. *Успехи современной биологии*. 2017;137(3):237–246.
[Mikhaylova Yu.V., Terent'eva L.Yu. Huge Mitochondrial Genomes in Embryophyta Plants. *Uspehi sovremennoy biologii = Advances in modern biology*. 2017;137(3):237–246 (in Russian)]
2. Kozik A. et al. The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genet*. 2019;15(8):e1008373.
3. Liu H. et al. Structural variation of mitochondrial genomes sheds light on evolutionary history of soybeans. *The Plant Journal*. 2021;108:1456–1472.
4. Chang S. et al. The Mitochondrial Genome of Soybean Reveals Complex Genome Structures and Gene Evolution at Intercellular and Phylogenetic Levels. *PLoS ONE*. 2013;8(2):e56502.
5. Shen Y. et al. Update soybean Zhonghuang 13 genome to a golden reference. *Sci China Life Sci*. 2019;62(9):1257–1260.
6. Handa H., Itani K., Sato H. Structural features and expression analysis of a linear mitochondrial plasmid in rapeseed (*Brassica napus* L.). *Mol Genet Genomics*. 2002;267:797–805.