БИОФИЗИКА, 2022, том 67, № 2, с. 234–239

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 519.876.5

АНАЛИЗ СТРУКТУРНОЙ ВАРИАБЕЛЬНОСТИ ГЕНОМОВ ЛЬНА *Linum usitatissimum* L.

© 2022 г. М.А. Дук*, А.А. Канапин*, А.А. Самсонова*, Т.А. Рожмина, М.Г. Самсонова***

**Санкт-Петербургский политехнический университет Петра Великого, 195251, Санкт-Петербург, Политехническая ул., 29*

***Институт льна – обособленное подразделение Федерального научного центра лубяных культур, 172002, Торжок Тверской области, ул. Луначарского, 35*

E-mail: m.samsonova@spbstu.ru

Поступила в редакцию 07.12.2021 г.
После доработки 07.12.2021 г.
Принята к публикации 12.12.2021 г.

У 100 хорошо секвенированных геномов льна проанализированы два типа структурной изменчивости: вариация присутствия/отсутствия генов и вариация числа копий. Показано, что в геномах рассматриваемых образцов льна в сравнении с референсным геномом наблюдаются делеции последовательностей (участков ДНК), вставки новых последовательностей и увеличение числа копий последовательностей. Функциональная аннотация соответствующих районов в референсном геноме и новых последовательностей показала, что они кодируют белки, участвующие в ответе растения на биотический и абиотический стрессы, в энергетическом обмене, вирусной и транспозонной активности, а также в формировании клеточных мембран. Выявленные функции могут свидетельствовать об адаптации сортов к региональным условиям выращивания посредством структурной изменчивости.

Ключевые слова: лен, геномика, секвенирование, структурная вариабельность

DOI: 10.31857/S0006302922020041

Лен является важной сельскохозяйственной культурой двойного назначения. Семена масличного льна являются ценным источником высококачественных ненасыщенных кислот, лигнанов, легко усваиваемых протеинов, диетической клетчатки, витаминов и минеральных элементов. Лен-долгунец служит основным источником натурального волокнистого сырья, в котором в настоящее время нуждается не только текстильная, но и другие высокотехнологичные отрасли экономики – фармацевтическая промышленность, космос, оборонный комплекс, автомобилестроение [1]. В современных условиях лен-долгунец рассматривается как стратегическая культура России, позволяющая заменить хлопок-сырец, который перешел в разряд импортного сырья.

Детальная характеристика генетического разнообразия льна имеет первостепенное значение для долгосрочной устойчивости и диверсификации производства этой сельскохозяйственной культуры, а также для общего успеха селекционных программ. В последнее время в этой области был достигнут значительный прогресс, в первую очередь благодаря публикациям результатов исследований генетического разнообразия льна из ряда национальных коллекций [2–6]. Коллекция льна, созданная в Федеральном центре лубяных культур (ФЦЛК), является одной из крупнейших в мире и охватывает практически все генетическое разнообразие этой культуры. Помимо современных отечественных и зарубежных сортов ФЦЛК располагает образцами семян ценных селекционных линий, староместных и кряжевых форм, а также дикорастущих видов, большинство из которых уже невозможно обнаружить в природе. Особенно важно подчеркнуть, что коллекция ФЦЛК включает сорта льна из Евразии с большой долей унаследованных русских местных форм, что отличает ее от коллекций, использованных в предыдущих генетических исследованиях.

Ранее мы охарактеризовали генетическое разнообразие в форме однонуклеотидных полиморфизмов у репрезентативной выборки образцов льна из коллекции ФЦЛК [7]. Мы наблюдали значительную дифференциацию популяций мас-

личного льна и льна-долгунца, идентифицировали области генома, маркированные сигналами недавней селекции, и показали, что они заметно отличаются у долгунцов и масличных форм, впервые попытались всесторонне охарактеризовать кряжи – староместные сорта русского происхождения, чтобы пролить свет на их селекционную историю и их связь с современными сортами льна-долгунца. Здесь мы приводим результаты анализа структурной вариабельности геномов образцов коллекции.

МАТЕРИАЛЫ И МЕТОДЫ

Коллекция из 100 образцов льна была выращена на опытном поле ФЦЛК в Торжке (Тверская обл.) в нее вошли 47 долгунцов, 24 межеумка, 10 крупносемянных образцов и 22 кудряша. Среди образцов обеих групп были представлены местные формы (ландрасы), кряжи (староместные сорта, выведенные российскими крестьянами в XIX веке), современные селекционные сорта и селекционные линии из 30 стран со всех континентов. ДНК из листьев, собранных у образцов, выделяли с помощью набора DNeasy Plant Mini (Qiagen, США).

Секвенирование ДНК было выполнено в BGI с использованием протокола Illumina, генерирующего считывания парных концов длиной 150 п.н. Было получено 9220.83 Гб необработанных данных, содержащих 6147221648 прочтений со средним покрытием 20.6×. Обработанные чтения были выровнены относительно NCBI-сборки референсного генома льна ASM22429v2 с помощью bwa-mem с использованием стандартных параметров [8].

Были проанализированы два типа структурной вариабельности – вариации присутствия/отсутствия генов и участков ДНК, а также вариации числа копий. При анализе вариаций присутствия/отсутствия генов вставки и делеции, наблюдаемые в геномах образцов, были проанализированы отдельно. Для анализа вставок с помощью SAMtools [9] были выбраны прочтения, не выравнивавшиеся на референсный геном льна, и собраны в длинные контиги с помощью программы ABySS [10]. С помощью алгоритма blat [11] контиги были проанализированы на степень совпадения с референсным геномом, для дальнейшего анализа в качестве новых вставок были выбраны прочтения, длина которых составляет более 1000 пар оснований, а совпадение с референсным геномом менее 25%. В выбранных прочтениях были найдены рамки считывания (ORFs), найденные домены сравнивали с актуальной базой данных Pfam [12], то же самое делалось для комплементарных последовательностей выбранных прочтений.

Вариации числа копий генов были найдены с помощью программы CNVnator [13]. Для анализа делеций и увеличения числа копий генов были найдены пересечения их с известными генами льна, а также на основе референсного генома выбраны участки, отсутствующие в образцах; домены, попавшие в эти участки, сравнивали с актуальной базой данных Pfam.

При сравнении с базой Pfam было обнаружено множество доменов, связанных с активностью транспозонов в геноме, которые присутствовали одновременно и среди вставок, и среди делеций, иными словами, не являлись новыми вставками, поэтому для анализа новых вставок подобные домены были исключены из рассмотрения для каждого образца. Для обобщения функций доменов, найденных среди вставок и делеций, было проведено сравнение с базой данных GO (Gene Ontology).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Общее число обнаруженных структурных вариантов в 100 образцах льна составило 216863, среди этих вариантов преобладали делеции (209294).

Анализ вариации присутствия/отсутствия. Суммарная длина контигов, не выравнивавшихся на референсный геном, составляла в среднем 8727963.5 п.н. при длине референсного генома 316167074 п.н., т. е. в среднем составляла порядка 3% генома у рассматриваемых образцов.

Функциональная аннотация последовательностей, отсутствующих в референсном геноме, но присутствуюших в индивидуальных образцах («новые вставки») с помошью базы данных Pfam выявила 1786 уникальных белковых доменов на прямой цепи ДНК и 623 на комплементарной цепи. В среднем каждый образец содержал 352 таких домена на прямой цепи ДНК и 74 домена на комплементарной цепи. На рис. 1 показано, какие белковые домены встречаются наиболее часто в среднем на каждый образец. Среди них можно отметить домены белков с антимикробными функциями (MatE, Antimicrobial10); анкириновые повторы, связанные со множеством функций (Ank, Ank_2, Ank_3, Ank_4, Ank_5); домены, необходимые для расщепления фитопатогенов (TAXi_C, TAXi_N); домены, связанные с энергетическим обменом (Polysacc_synt_C, Glyco_transf_61, PC_Esterase); домены вирусных белков (Luteo_Vpg).

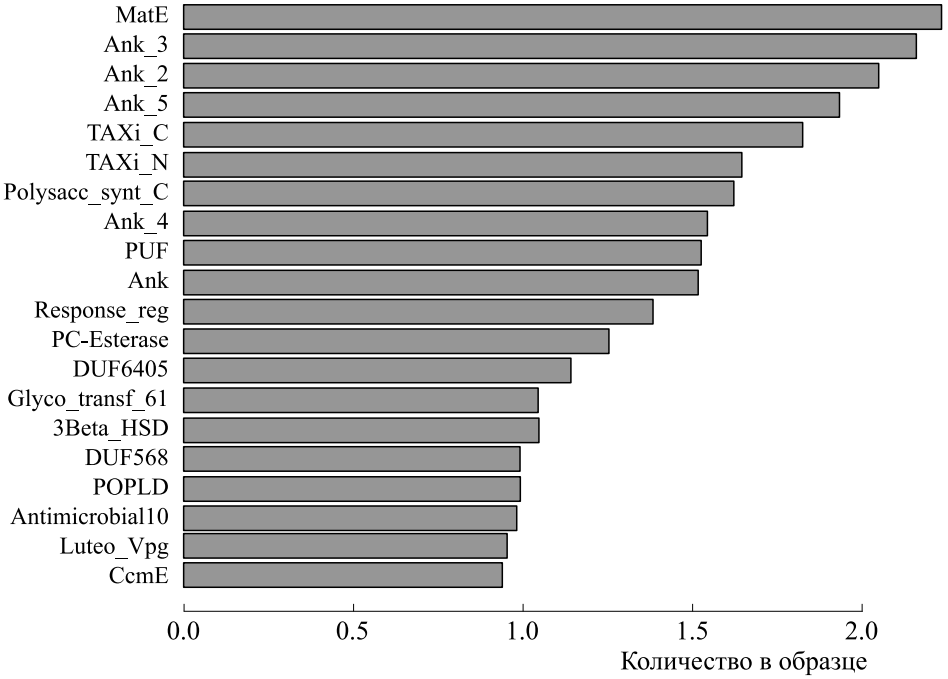

Рис. 1. Функциональная аннотация и среднее число на образец белковых доменов, кодируемых «новыми вставками» последовательностей.

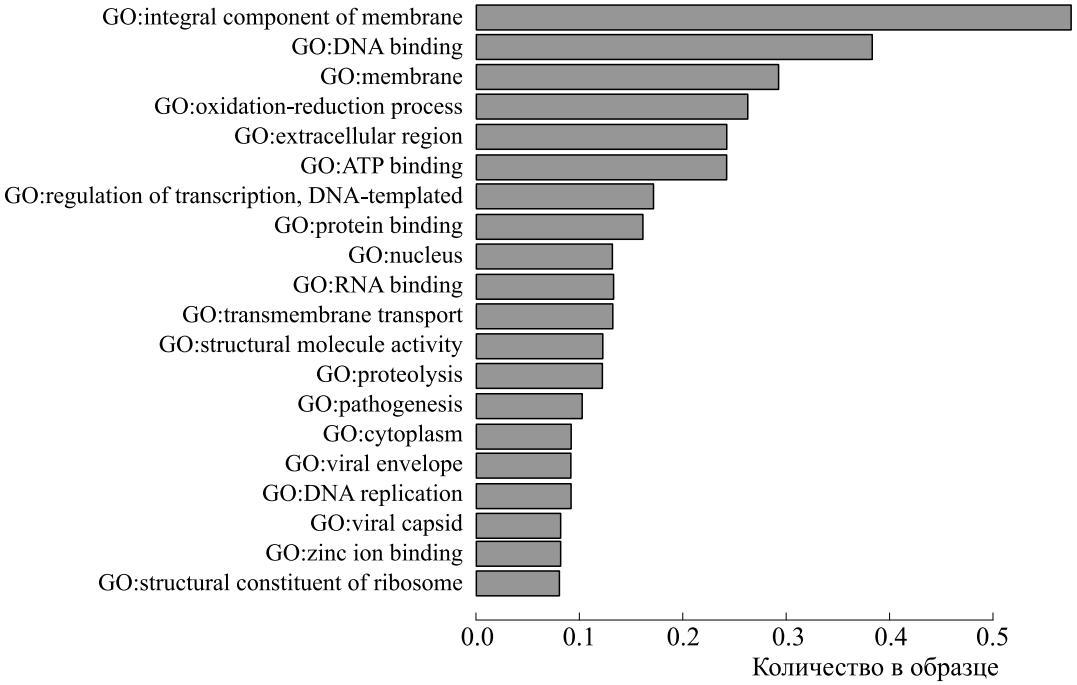

Рис. 2. Функциональная аннотация (по Gene Onthology) всех найденных доменов в новых вставках последовательностей и их среднее число на образец.

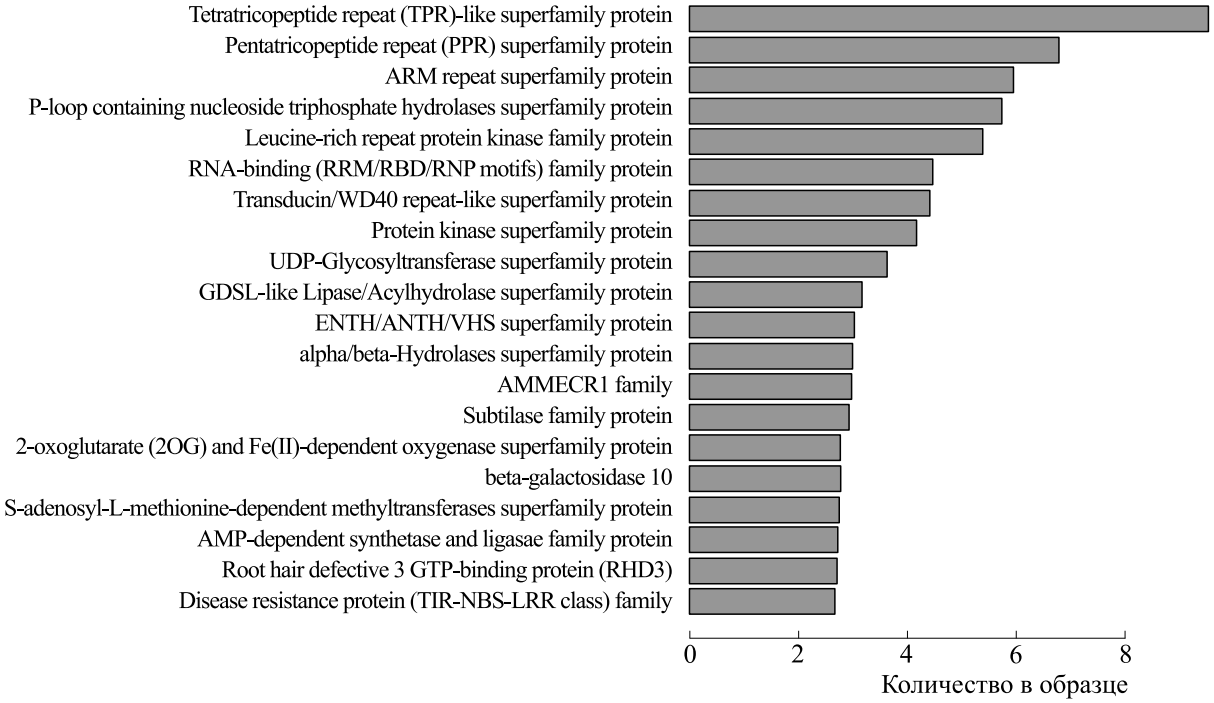

Рис. 3. Гены, наиболее часто пересекающиеся с делециями, объединенные по функциям.

На рис. 2 показан график, обобщающий с помощью GO (Gene Onthology) функции всех найденных доменов в новых вставках последовательностей в геномах образцов. Следует отметить, что наиболее часто встречающиеся домены, показанные ранее на рис. 1, не аннотированы по GO и на рис. 2 не представлены, однако по доменам, аннотированным по GO, также можно сделать вывод, что вставки (т. е. последовательности, присутствующие в образцах, но отсутствующие в референсе) часто кодируют белки, связанные с ДНК-процессами, энергетическим обменом, прочностью клеточных мембран, вирусной активностью.

Анализ вариации числа копий. Случаи уменьшения числа копий (т. е. делеции) и увеличения числа копий были проанализированы по отдельности. При анализе делеций были проанализированы гены, попадающие в такие районы. На рис. 3 показан обобщенный график функций генов, наиболее часто пересекающихся с выявленными делециями у рассматриваемых образцов. Здесь можно отметить белки, связанные с ответом на окислительный и солевой стресс и иммунитетом (TPR, PPR, ARM); белки, связанные с реакцией на засуху и колебания температуры (WD40, TPR, LRR); различные ферменты (протеинкиназы, гидролазы, гликозилтрансферазы, метилтрансферазы), связанные с энергетическим обменом. При анализе увеличения числа копий генов в среднем обнаруживалось 76 подобных случаев на образец. Аннотация таких участков показала, что они содержат гены, связанные с реакцией на абиотический стресс (RING/U-box, PPR, TTR, LRR), различные ферменты, белки, связанные с формированием семян и реакцией на освещенность (рис. 4).

Подводя итог анализу вариаций присутствия/отсутствия генов и участков ДНК и вариаций числа копий, можно отметить, что в геномах рассматриваемых образцов льна в сравнении с референсным геномом наблюдаются делеции последовательностей (участков ДНК), вставки новых последовательностей и увеличение числа копий последовательностей. Функциональная аннотация соответствующих районов в референсном геноме и новых последовательностей показала, что они кодируют белки, участвующие в ответе растения на биотический и абиотический стрессы, в энергетическом обмене, вирусной и транспозонной активности, а также в формировании клеточных мембран. Выявленные функции могут свидетельствовать об адаптации сортов к региональным условиям выращивания посредством структурной изменчивости генома.

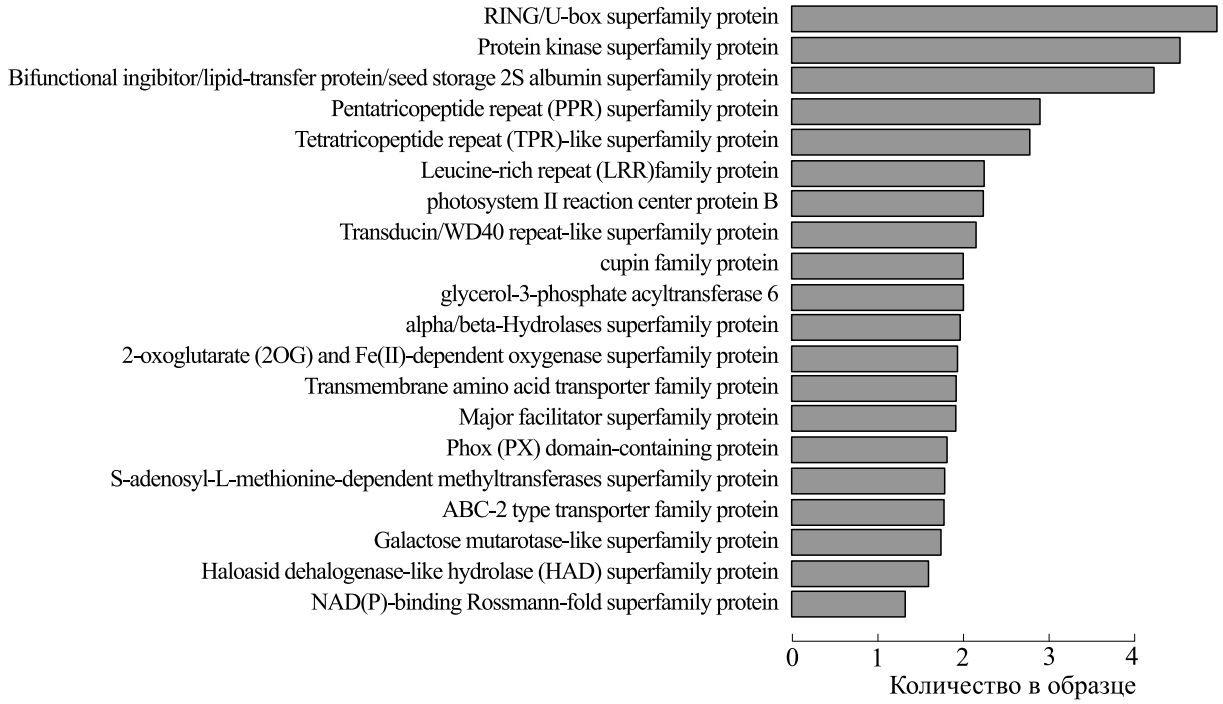

Рис. 4. Гены, наиболее часто пересекающиеся с увеличением числа копий, объединенные по функциям, и их среднее количество в образце.

ФИНАНСИРОВАНИЕ РАБОТЫ

Исследование выполнено при финансовой поддержке Российского научного фонда (проект № 19-16-00030).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. Т. А. Рожмина, Л. Н. Павлова, Н. В. Мельникова и Л. М. Голубева, Успехи соврем. науки **1** (10), 184 (2017).
2. A. Diederichsen, P. M. Kusters, D. Kessler, et al., Gen. Resources Crop Evol. **60**, 1479 (2012). DOI: 10.1007/s10722-012-9936-1
3. B. J. Soto-Cerda, A. Diederichsen, R. Ragupathy, and S. Cloutier, BMC Plant Biol. **13**, 78, (2013). DOI: 10.1186/1471-2229-13-78
4. F. M. You, J. Xiao, P. Li, et al., Int. J. Mol. Sci. **19** (8), 2303 (2018). DOI: 10.3390/ijms19082303
5. D. Guo, H. Jiang, W. Yan, et al., Front. Plant Sci. **10**, 1682 (2019). DOI: 10.3389/fpls.2019.01682
6. Chandrawati, N. Singh, R. Kumar, et al., Physiol. Mol. Biol. Plants **23**, 207 (2017). DOI: 10.1007/s12298-016-0408-5
7. M. Duk, A. Kanapin, S. Surkova, et al., Front. Plant Sci. **12**, 764612 (2021). DOI: 10.3389/fpls.2021.764612
8. H. Li and R. Durbin, Bioinformatics **25**, 1754 (2009).
9. P. Danecek, J. K. Bonfield, J. Liddle, et al., Gigascience **10** (2), giab008 (2021). DOI: 10.1093/gigascience/giab008
10. G. Robertson, J. Schein, R. Chiu, et al., Nat. Methods **7** (11), 909 (2010). DOI: 10.1038/nmeth.1517
11. W. J. Kent, Genome Res. **12** (4), 656 (2002).
12. J. Mistry, S. Chuguransky, L. Williams, et al., Nucl. Acids Res. **49** (D1), D412 (2021). DOI: 10.1093/nar/gkaa913
13. A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein, Genome Res. **21** (6), 974 (2011). DOI: 10.1101/gr.114876.110

Analysis of Structural Variation in the Genome of Flax *Linum usitatissimum* L.

M.A. Duk*, A.A. Kanapin*, A.A. Samsonova*, T.A. Rozhmina, and M.G. Samsonova***

**Peter the Great St. Petersburg Polytechnic University, ul. Polytekhnicheskaya 29, St. Petersburg, 195251 Russia*

***Flax Institute, ul. Lunacharskogo 35, Thorzhok, Tver Region, 172002 Russia*

Two types of structural variation such as presence-absence and copy number variations were analyzed in 100 well-sequenced flax genomes. In this study, we observed deletions of DNA sequences, insertions of new sequences, and copy number amplification in individual flax varieties compared to the reference genome. The functional annotation of the corresponding regions in the reference genome and new sequences showed that they encode proteins involved in the plant response to biotic and abiotic stresses, in energy metabolism, viral and transposon activity, and in the formation of cell membranes. Our analysis demonstrates that identified functions might be indicative of adaptation of varieties to regional growing conditions through structural variation.

Keywords: flax, genomics, sequencing, structural variation