БИОФИЗИКА, 2023, том 68, № 3, с. 501–505

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 519.876.5

ГЕНЕТИЧЕСКИЕ ДЕТЕРМИНАНТЫ ЦЕЛОСТНОСТИ ГЕНОМА ЛЬНА

© 2023 г. А.А. Канапин*, А.А. Самсонова*,#

**Санкт-Петербургскии государственный университет, Университетская набережная, 7–9, Санкт-Петербург, 199034, Россия*

#E-mail: a.samsonova@spbu.ru

Поступила в редакцию 06.12.2022 г.
После доработки 06.12.2022 г.
Принята к публикации 21.12.2022 г.

Последние достижения в области методов высокопроизводительного секвенирования позволили разработать инновационный подход к оценке стабильности и целостности генома. Глубина сигнала покрытия в определенной точке генома может указывать на потерю целостности ДНК в регионе. В данной работе мы превратили ранее разработанную метрику локальной целостности генома, оценивающую равномерность сигнала покрытия, в количественный признак и провели поиск связанных с ним генетических вариантов в геноме льна. Другими словами, мы применили методологию *x*QTL (т.е. *x* Quantitiave Trait Loci, где *x* – обозначение произвольной количественной характеристики, связанной с определенным регионом генома, например, уровень экспрессии генов, степень покрытия рибосомами и т.д.) для выявления геномных регионов, вероятно способствующих потере целостности генома и, возможно, участвующих в поддержании стабильности генома. Анализ, проведённый с использованием данных полногеномного секвенирования 100 образцов льна, позволил идентифицировать гены, вероятно, принимающие участие в поддержании стабильности генома у льна и, возможно, в целом у растений, а также обозначить новые процессы, связанные с данным процессом.

Ключевые слова: стабильность генома, целостность генома, полногеномное секвенирование, локусы количественных признаков, геномика растений, лен, Linum usitasissimum L.

DOI: 10.31857/S0006302923030110, **EDN:** FRUWBQ

Поддержание стабильности и целостности генома является одним из наиболее важных клеточных процессов, обеспечивающих нормальное функционирование организма и передачу наследственных признаков в ряду поколений. Нарушения в работе этих механизмов приводят к повышению уровня геномных аберраций, таких как хромосомные перестройки, варианты копийности и другие структурные вариации. Данные процессы достаточно хорошо изучены у животных и человека, особенно в связи с патологическими последствиями, такими как наследственные и онкологические заболевания, возникающие в результате нарушения нормального функционирования механизмов защиты генома и поддержания его целостности [1–7].

У растений роль структурных вариаций, возникающих как следствие нарушения целостности генома, высока в силу масштабности вносимых ими изменений в геном растения [1, 2, 8]. Изменения эпигенетического ландшафта, нарушения механизмов упаковки хроматина, репликации и трансляции ДНК могут приводить к тому, что те или иные участки генома с большей или меньшей вероятностью подвергнутся изменениям и приобретут структурные варианты. Методы секвенирования нового поколения дают возможность получить прямую или косвенную информацию о состоянии различных геномных регионов. В частности, данные о глубине покрытия секвенирования позволяют предсказывать структурные варианты и варианты копийности [9, 10]. В то же время, использование сигнала покрытия, а именно информации о количестве фрагментов секвенирования, выравненных на референсный геном в той или иной его точке может служить индикатором целостности генома. Количество фрагментов секвенирования, картированных на тот или иной локус генома, зависит от концентрации геномной ДНК ему соответствующей, и, следовательно может варьироваться при несбалансированных хромосомных перестройках, таких как делеции и дупликации. Ранее нами был разработан прототип количественной метрики локальной целостности генома, основанный на анализе свойств сигнала глубины покрытия секвенирования [11]. В работе, результаты которой представлены ниже, метрика используется в качестве ко-

личественного признака для анализа ассоциаций типа *x*QTL (Quantitative Trait Loci, локусы количественного признака) [12–14], что позволило идентифицировать гены, вероятно участвующие в поддержании целостности генома у растений.

МАТЕРИАЛЫ И МЕТОДЫ

В качестве исходных данных были использованы результаты секвенирования 100 образцов льна (*Linum usitasissimum* L.), секвенированных по протоколу Illumina парными прочтениями длиной по 100 п.о., средняя глубина покрытия секвенирования составляла 20×. Образцы льна были получены из коллекции Федерального научного центра лубяных культур (Институт льна, Торжок, Россия) [15]. Выравнивание фрагментов секвенирования на геном льна версии GCA_000224295.2_ASM22429v2 проводили при помощи программы bwa mem [16]. Поиск однонуклеотидных замен проводили с помощью программы NGSEP, версия 4.0.0 [17]. Предварительная фильтрация найденных вариантов с порогами по следующим параметрам: MAF > 0.05 (Minor Allele Frequency, частота минорного аллеля), частота встречаемости варианта – не ниже 0.85. Общее число вариантов, использованных для анализа, составляло 1873082 после фильтрации с указанными параметрами.

Вычисление метрики локальной стабильности генома проводили по ранее описанной методике [11] в регионах длиной 16384 п.о. на всех 15 хромосомах. Для более точного анализа ассоциаций были рассчитаны значения 10 ковариат при помощи пакета PEER, версия 1.3 [18]. Значения метрики локальной стабильности генома использовали в качестве количественного признака в анализе *x*QTL (quantitative trait loci, локусы количественного признака). Для поиска *цис*- и *транс*-вариантов, ассоциированных со значениями метрики, применяли пакет QTLtools, версия 1.3.1 [19]. Пороговое значение параметра FDR (false discovery rate, доля ложноположительных значений) для фильтрации обнаруженных вариантов составляло 0.01. Общее число однонуклеотидных вариантов, обнаруженных в данном анализе, составило 4630; из них для более детального рассмотрения были отобраны только те, которые попадали в кодирующие участки генома – общим числом 1020 вариантов в 947 генах. Анализ обогащенности для категорий GO (Gene Ontology), соответствующих данным генам, проводили в пакете программ XGR [20]. Функциональную аннотацию белков льна проводили при помощи поиска гомологов с белками *Arabidopsis thaliana* и идентификации функциональных доменов, представленных в базе данных Pfam [21].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Анализ метрики локальной стабильности генома как количественного признака позволил идентифицировать 4481 однонуклеотидный вариант, ассоциированный с ней. С целью функциональной аннотации генов, содержащих обнаруженные варианты (общим числом 923), мы провели анализ обогащенности категориями Gene Ontology (GO), результаты приведены на рис. 1.

Наличие таких категорий, как GO:0003678 – DNA helicase activity и GO:0006260 – DNA replication указывает на участие белков, принадлежащих к данным категориям, в процессах, связанных с функционированием генетического аппарата клетки и, следовательно, имеющих прямое отношение к стабильности генома. В частности, следует отметить такие гены, как *Lus10036982* – DNA polymerase alpha 2, *Lus10005256* и *Lus10016124* – AAA-АТФазы, *Lus10023341* – DNA repair (Rad51) family protein, *Lus10020209* – ankyrin repeat family protein, *Lus10024204* – basic leucine-zipper, *Lus10030531* – Basic-leucine zipper (bZIP) transcription factor family protein, *Lus10024246* – WRKY DNA-binding protein, *Lus10013294* – DEAD/DEAH box helicase.

Значительная часть белков также способна связываться с АТФ и АДФ, что следует из анализа сверхпредставленности доменов Pfam (см. табл. 1)

Значительная доля белков включает домены типа TIR (Toll- interleukin receptor), выполняющие сигнальную рецепторную функцию, ассоциированную с врожденным иммунитетом у животных и, вероятно, иммунным ответом у растений [22, 23]. Однако роль доменов такого типа в растениях, в частности, в процессах, обеспечивающих устойчивость к патогенам, еще остается не конца изученной.

Большинство генетических вариантов, ассоциированных с метрикой локальной целостности генома и обладающих способностью связываться с АТФ и АДФ принадлежат к генам двух типов – Disease resistance protein (TIR-NBS-LRR class) и NAD(P)-binding Rossmann-fold superfamily protein (рис. 2).

Эти белки содержат домены, связывающиеся с нуклеиновыми кислотами и нуклеотидами, такие как NB-ARC [24], Rossmann-fold [25], P-loop АТФ/ГТФазы [26] и играют важную роль в процессах, связанных с устойчивостью к патогенам. Однако другие их функции все еще остаются мало исследованными. В частности, белки, относящиеся к классу TIR-NBS-LRR являются активаторами многих сигнальных каскадов, модифицируя различные типы протеинкиназ. Таким образом, они могут оказывать прямое влияние на поддержание стабильности генома за счет реакции на изменения внешней и внутренней среды клетки.

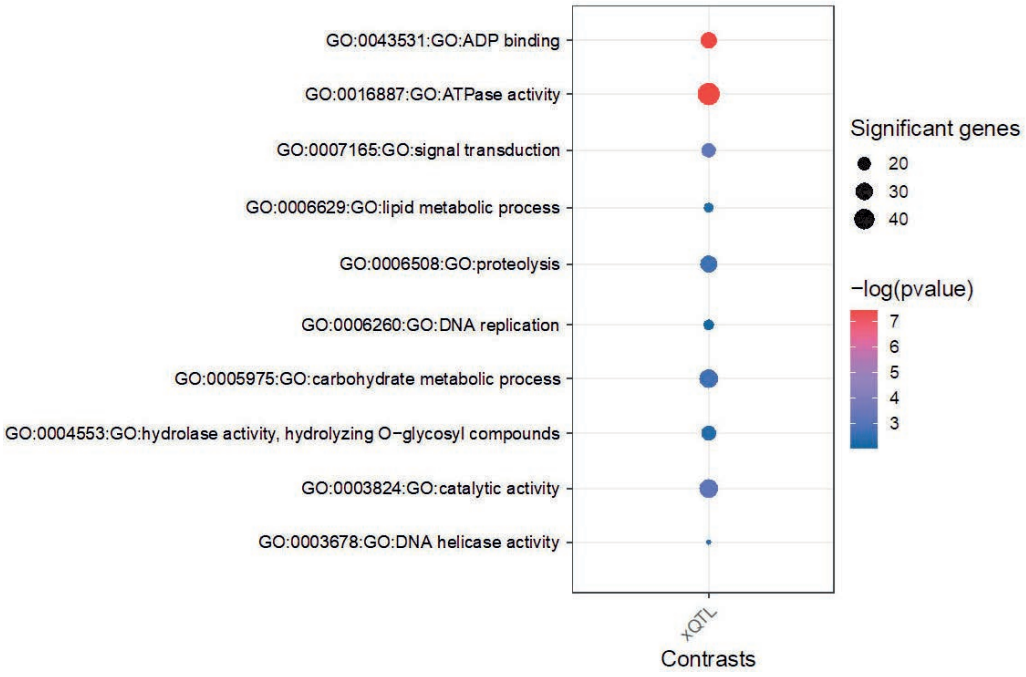

Рис. 1. Категории Gene Ontology, сверхпредставленные в генах, несущих варианты, ассоциированные с метрикой стабильности генома. Размер точки пропорционален количеству генов, аннотированных данной категорией GO, интенсивность окраски соответствует обратному логарифму *p*-value.

Можно предположить, что АТФазная активность белков, кодируемых идентифицированными генами, в той или иной форме связана с затратами энергии АТФ на процессы поддержания и восстановления стабильности генома. Поскольку многие процессы защиты растений от патогенов связаны с модификацией эпигенетического ландшафта, возможно, что сходные процессы также ассоциированы с поддержанием целостности генома на разных уровнях.

Таблица 1. Домены Pfam, сверхпредставленные в генах, ассоциированных с метрикой стабильности генома

PFAM ID	PFAM name	Число генов	*p-value*
PF01582	TIR	21	4.90E-09
PF13676	TIR_2	21	4.90E-09
PF00004	AAA	35	6.20E-08
PF00931	NB-ARC	27	8.50E-08
PF13191	AAA_16	44	1.40E-07
PF13401	AAA_22	39	3.30E-06
PF05729	NACHT	25	3.50E-06
PF13855	LRR_8	44	6.20E-06
PF01637	ATPase_2	17	7.50E-06
PF12799	LRR_4	46	9.60E-06
PF00432	Prenyltrans	8	1.80E-05

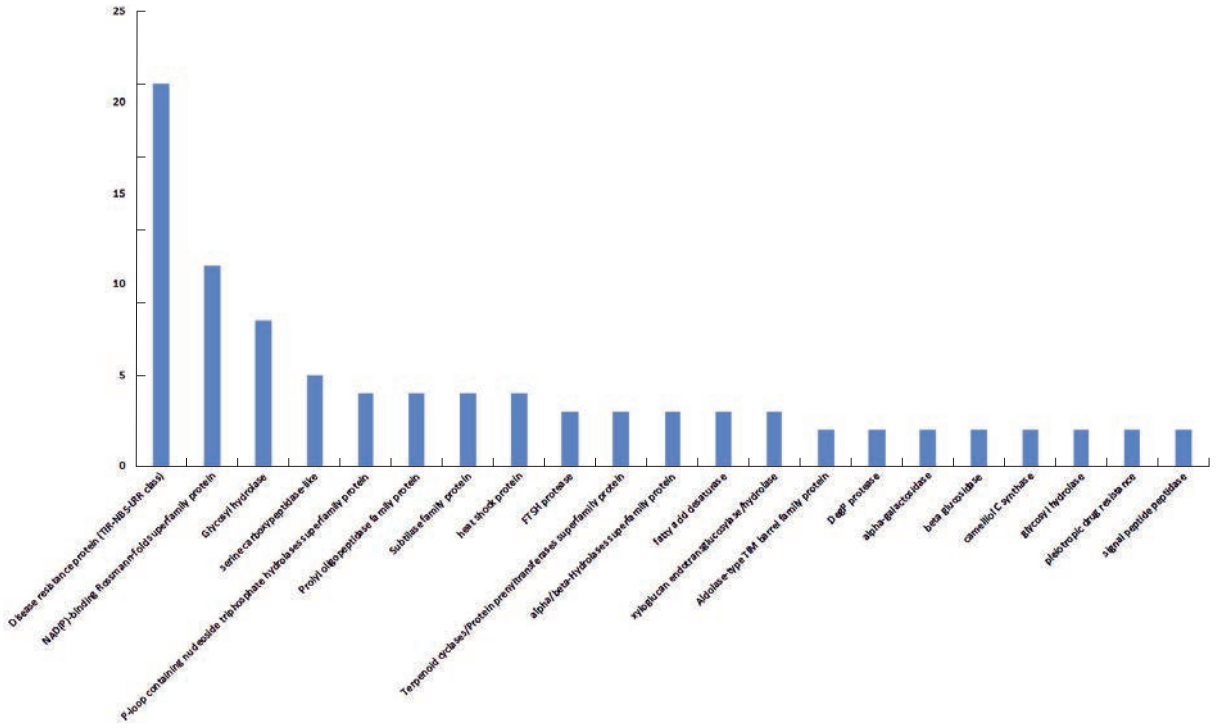

Рис. 2. Распределение генов по функциональной аннотации.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при финансовой поддержке Российского научного фонда (грант № 20-14-00072).

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. Z. N. Lye, M. D. Purugganan, Trends Plant Sci., **24**, 352 (2019).
2. Y. Yuan, P. E. Bayer, J. Batley, and D. Edwards, Plant Biotechnol. J., **19**, 2153 (2021).
3. S. Nik-Zainal, Genome Med., **11**, 1 (2019).
4. A. Janssen, S. U. Colmenares, and G. H. Karpen, Annu. Rev. Cell Dev. Biol., **34**, 265 (2018).
5. S. S. Ho, A. E. Urban, and R. E. Mills, Nat. Rev. Genet., 1 (2019).
6. E. M. Kass, M. E. Moynahan, and M. Jasin, MOLCEL, **62**, 777 (2016).
7. N. Andor, C. C. Maley, and H. P. Ji, Cancer Res., **77**, 2179 (2017).
8. A. Dolatabadian, D. A. Patel, D. Edwards, and J. Batley, Theor. Appl. Genet., **130**, 2479 (2017).
9. A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein, Genome Res., **21**, 974 (2011).
10. V. Boeva, et al., Bioinformatics, **28**, 423 (2011).
11. A. Samsonova, et al., Int. J. Mol. Sci., **22**, 2665 (2021).
12. D. Arends, et al., Bioinformatics, **28**, 1042 (2012).
13. B. Ng, et al., Nat. Neurosci., **20**, 1 (2017).
14. Y. Ma, H. Klein, and P. L. D. Jager, Brain Pathol., **30**, 984 (2020).
15. 15. М. А. Дук, А. А. Канапин, А. А. Самсонова и др., Биофизика, **67**, 234 (2022).
16. H. Li and R. Durbin, Bioinformatics, **26**, 589 (2010).
17. D. Tello, et al., Bioinformatics, **35**, 4716 (2019).
18. O. Stegle, L. Parts, M. Piipari, et al., Nat. Protoc., **7**, 500 (2012).
19. H. Ongen, et al., Nat. Commun., **8**, 1 (2017).
20. H. Fang, B. Knezevic, K. L. Burnham, and J. C. Knight, Genome Med., **8**, 1 (2016).
21. J. Mistry, et al., Nucl. Acids Res., **49**, gkaa913 (2020).
22. J.-M. Zhou and Y. Zhang, Cell, **181**, 978 (2020).
23. D. Lapin, O. Johanndrees, Z. Wu, et al., Plant Cell, **34**, 1479 (2022).
24. S. J. Riedl, W. Li, Y. Chao, et al., Nature, **434**, 926 (2005).
25. N. Boes, K. Schreiber, E. Härtig, et al., J. Bacteriol., **188**, 6529 (2006).
26. D. D. Leipe, Y. I. Wolf, E. V. Koonin, and L. Aravind, J. Mol. Biol., **317**, 41 (2002).

Genetic Determinants of Flax Genome Integrity

A. A. Kanapin* and A.A. Samsonova*

**St. Petersburg State University, Universitetskaya nab. 7–9, St. Petersburg, 199034 Russia*

Recent advances in high-throughput sequencing methods have enabled development of an innovative approach to evaluation of genome stability and integrity. The depth of the coverage signal at a particular location of the genome may indicate the loss of DNA integrity in the region. In this work, the previously developed metric of local genome integrity that estimates the uniformity of coverage signal is considered a quantitative trait and a search for genetic variants associated with the uniformity of coverage signal in flax genome is performed. In particular, quantitative trait locus (*x*QTL) analyses (i.e., *x* Quantitiave Trait Loci, where *x* is the designation of an arbitrary quantitative characteristic associated with a particular genome region; for example, the level of gene expression, the degree of ribosome coverage, etc.) have been applied to identify genomic regions that most likely contribute to loss of genome integrity and are, probably, involved in the maintenance of genome stability. The analysis carried out using information on whole-genome sequence assembly of 100 flax samples enabled identification of genes potentially implicated in genome integrity maintenance in flax and, possibly, in plants in general and also revealed novel processes associated with the maintenance of genome integrity.

Keywords: genome stability, genome integrity, whole genome sequencing, quantitative trait loci, plant genomics, flax, Linum usitasissimum L.