

Научная статья

УДК 575.224.22

DOI: <https://doi.org/10.21285/2227-2925-2022-12-4-599-604>



Биоинформатический метод определения однонуклеотидных полиморфизмов на примере гена *WIN* у *Glycine max*

Павел Дмитриевич Тимкин, Андрей Андреевич Пензин

Всероссийский научно-исследовательский институт сои,

г. Благовещенск, Российская Федерация

Автор, ответственный за переписку: Тимкин Павел Дмитриевич, tpd@vniisoi.ru

Аннотация. В данной работе предлагается гипотетический метод поиска SNPs (однонуклеотидных полиморфизмов) на примере гена рибонуклеазы *WIN*. Рибонуклеаза – фермент, который участвует в реакциях защиты против грибковых инфекций у сои, а также в других реакциях по защите от биотического стресса. Принадлежность к группе РНК-аз обуславливает ее специфические свойства, а именно способность к деградации чужеродных нуклеиновых кислот. Данная способность позволяет индуцировать общий неспецифический иммунный ответ растения на вторжение антигенных структур. В современной биотехнологии стоит задача разработки молекулярных методов и подходов, которые позволят увеличить резистентные свойства культуры или ускорить процессы ее адаптации в полевых условиях. Ключом к решению этой задачи может служить использование технологий искусственного индуцирования однонуклеотидных полиморфизмов в тех участках генома, которые кодируют белки, способные принимать участие в защитных реакциях против биотического стресса. В ходе проведенного исследования нам удалось предположить 5 однонуклеотидных полиморфизмов с использованием методов биоинформационного анализа в рамках описываемой нами методики. Локализация и детекция SNPs является сложной задачей ввиду наличия изменения одного нуклеотида. Поэтому в практике биотехнологов имеется задача проведения предиктивного анализа с целью локализации потенциальной последовательности нахождения однонуклеотидного полиморфизма. После выяснения гипотетического расположения SNPs появляется возможность дальнейшего их детектирования с использованием сложных молекулярных методов, таких как ПЦР в реальном времени или локальное секвенирование. Данная технология даст возможность исследователям получить мощный инструмент для проведения селекционных работ с целью выведения сортов сои с заранее заданными свойствами. Подобные теоретические и предсказательные модели позволят более оперативно реагировать на изменяющуюся обстановку в условиях антропогенной и техногенной нагрузки на экологическую растительную среду.

Ключевые слова: *WIN*, *Glycine max*, SNPs-маркеры, *in silico*, селекция, биоинформатика

Для цитирования: Тимкин П. Д., Пензин А. А. Биоинформатический метод определения однонуклеотидных полиморфизмов на примере гена *WIN* у *Glycine max* // Известия вузов. Прикладная химия и биотехнология. 2022. Т. 12. N 4. С. 599–604. <https://doi.org/10.21285/2227-2925-2022-12-4-599-604>.

PHYSICOCHEMICAL BIOLOGY

Original article

Bioinformatic method for determining single nucleotide polymorphisms on the example of gene *WIN* in *Glycine max*

Pavel D. Timkin, Andrey A. Penzin

All-Russian Research Institute of soybean, Blagoveshchensk, Russian Federation

Corresponding author: Pavel D. Timkin, tpd@vniisoi.ru

Abstract. In this paper, a hypothetical method for locating SNPs (single nucleotide polymorphisms) on the example of the ribonuclease gene *WIN* was proposed. Ribonuclease comprises an enzyme that participates in defence reactions against fungal infections in soybeans, as well as other protective responses to biotic stress. Its belonging to the RNA-ases group determines the specific properties, namely the ability to degrade foreign nucleic acids. This ability provides for a general nonspecific immune response of the plant to the invasion of antigenic structures. Modern biotechnology calls for the development of molecular methods and approaches that will increase the resistance of a culture or accelerate the processes of its adaptation in

the field. This problem can be solved by using technologies of SNP artificial induction in those parts of the genome that encode proteins capable of acting in protective reactions against biotic stress. In the study, 5 single-nucleotide polymorphisms were proposed using bioinformatic analysis. Since the localisation and detection of SNPs comprise a challenging task due to the presence of a single nucleotide change, in the biotechnological practice, predictive analysis is carried out in order to localise the potential sequence of occurring single-nucleotide polymorphism. Following the identification of the hypothetical SNP location, they can be further detected using complex molecular methods, such as real-time PCR or local sequencing. This technology can become a powerful tool for breeding soybean varieties having predetermined properties. Such theoretical and predictive models will allow for a quicker response to the dynamic environment under manmade load on plants.

Keywords: WIN, *Glycine max*, SNPs markers, *in silico*, selection, bioinformatics

For citation: Timkin P. D., Penzin A. A. Bioinformatic method for determining single nucleotide polymorphisms on the example of gene WIN in *Glycine max*. *Izvestiya Vuzov. Prikladnaya Khimiya i Biotekhnologiya* = *Proceedings of Universities. Applied Chemistry and Biotechnology*. 2022;12(4):599-604. (In Russian). <https://doi.org/10.21285/2227-2925-2022-12-4-599-604>.

ВВЕДЕНИЕ

SNPs (однонуклеотидный полиморфизм) – это отличия последовательности ДНК размером в один нуклеотид (А, Т, G или С) в геноме (или в другой сравниваемой последовательности) представителей одного вида или между гомологичными участками гомологичных хромосом [1]. Замены нуклеотидов в гене бывают 2-х типов: синонимичные, когда замена одного нуклеотида в триплете не ведет к изменению аминокислоты в пептиде, и несинонимичные, когда изменение происходит.

Стоит отметить, что причины и механизмы, ведущие к появлению однонуклеотидных полиморфизмов, такие же, как у мутаций. К ним можно отнести мутации nonsense/missense. Missense-мутацией называют такую мутацию, в которой происходит замена нуклеотида, приводящая к замене аминокислоты в полипептиде [2]. В свою очередь, однонуклеотидная замена в nonsense-мутации приводит к образованию стоп-кодона. Номенклатурное различие крайне условно и формально, так, некоторые исследователи относят замену нуклеотида в гене к SNPs при условии, что данная замена встречается у более чем 1% от общей популяции [3].

Интерес к однонуклеотидным полиморфизмам у селекционеров и генетиков вызван тем, что их детекция и выявление свойств позволяет ускорить процесс создания новых сортов сои с заранее заданными свойствами, строить генетические карты, выявлять взаимосвязь фенотипа с генотипом [3, 4]. Внесение полиморфизмов в геном даст возможность повысить резистентность сортов сои к разнообразным биотическим или абиотическим стрессам. Также SNPs позволят задать нужные биохимические качества, например, повышенное содержание белка или липидов [5]. SNPs могут встречаться на протяжении всего генома как в его кодирующей, так и не кодирующей частях. В не кодирующей части наибольший интерес представляют так называемые QTL и SSR-маркеры, однако предсказание лока-

лизации их SNPs – более сложная и комплексная задача [6].

Современные методики белковой инженерии позволяют не проводить долгий мониторинг возможных нуклеотидных полиморфизмов и отбор нужных вариантов, после чего осуществляется гибридизация с целью закрепления полезных SNPs [7, 8]. С использованием биоинформатического анализа перед исследователями в области биотехнологии открываются возможности искусственной индукции подобных полиморфизмов с заранее заданными и известными свойствами. В литературе уже имеются описания применения подобных подходов в лабораторной практике [9]. Однако перед индукцией новых полиморфизмов требуется составление списка потенциальных кандидатов. Биоинформатика предоставляет широкий инструментарий для решения предсказательных задач в области прикладной генетики.

Целью данного исследования являлась разработка методологии предсказания локализации однонуклеотидных полиморфизмов в белок кодирующем гене *in silico* на основании информации об интересующем полипептиде.

В данной работе в качестве примерной модели предсказания был взят ген сои WIN, который согласно приведенным сведениям из базы данных Европейского института биоинформатики является фактором неспецифической резистентности к бактериальным или грибковым инфекциям^{1,2} [10]. Данное свойство может объясняться тем, что этот ген кодирует фермент рибонуклеазу, относящуюся к семейству РНК-аз, т. е. участвует в процессах деградации и распада нуклеиновых кислот в нормальном метаболизме самой клетки, а в случае инфекционных патологий выступает как часть неспецифического иммунного ответа.

Поиск информации из баз данных NCBI и Ensembl показал, что в данный момент не было обнаружено ни одного однонуклеотидного полимор-

¹Database for annotations of proteins by EMBL's EBI // European Molecular Biology Laboratory European Bioinformatics Institute [Электронный ресурс]. URL: <https://www.ebi.ac.uk/QuickGO/term/GO:0042742> (02.12.2022).

²Database for annotations of proteins by EMBL's EBI // European Molecular Biology Laboratory European Bioinformatics Institute [Электронный ресурс]. URL: <https://www.ebi.ac.uk/QuickGO/term/GO:0050832> (02.12.2022).

физма для гена *WIN* у *Glycine max*^{3,4}. Отсутствие информации об SNPs увеличивает спрос на поиск и локализацию этих маркеров для данного гена.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Исследования проводились на базе лаборатории биотехнологии Всероссийского научно-исследовательского института сои. Информация об исследуемых белках была взята из открытой базы Uniprot⁵. С целью выявления полиморфизмов (*WIN*) был проведен ряд действий, согласно разработанной методике (рис. 1). С использованием алгоритма BLAST при заданном родстве в 35–99% были найдены белки с максимально схожей последовательностью. На основании степени родства и наличия аннотаций в базе данных Европейского института биоинформатики (EMBL)⁶ был подобран максимально гомологичный белок. Затем было проведено выравнивание аминокислотных последовательностей, найдены консервативные участки. На заключительном этапе работы среди ранее найденных участков были отобраны те, что являются полиморфными.

Отбор участков первичной структуры белка, несущих в себе информацию о возможном однону-



Рис. 1. Блок-схема последовательных действий в разработанной методологии, используемой в данном исследовании

Fig. 1. Flowchart of sequential actions in the developed methodology used in the study

клеотидном полиморфизме в первичной структуре ДНК, производился вручную. Поиск осуществлялся среди всех найденных в процессе выравнивания консервативных участков полипептидов.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Во время идентификации целевой последовательности по базе Uniprot⁵ была взята полипептидная последовательность, кодируемая геном *WIN*, имеющая индекс по базе Q02243⁷. Данный фермент представляет собой короткую полипептидную цепь длиной в 102 аминокислоты и имеет молекулярную массу 11.124 kDa. После использования алгоритма BLAST был получен список из 250 полипептидов в ранге схожести от 35 до 99%, наиболее высокой степенью гомологии и наличием высокого кластера аннотаций обладал белок HEVL (рис. 2) ввиду высокой степени родства – 75%. HEVL относительно большой полипептид длиной 212 аминокислот. В базе данных он аннотируется с индексом P43082, является экспрессируемым у *Arabidopsis thaliana* (Резуховидка Таля) [12]. *Arabidopsis thaliana* – один из наиболее популярных объектов для генетических исследований ввиду короткого жизненного цикла, высокой плодовитости, малого размера генома, широкого ареала распространения [13–16]. Популярность для исследований данной культуры обеспечивает большой массив данных и аннотаций о геноме, протеоме и даже транскриптоме [17–19], что при высокой степени родства определенных генов и белков с *Glycine max* позволяет экстраполировать данные по гомологичному принципу.

На рис. 2 представлен финальный результат отбора с учетом ранее указанных метрических показателей. Такие метрики, как таксономическая принадлежность, длина первичной последовательности белка, наличие информации о 3-мерной структуре в PDB (Protein Data Bank) и т. д.,

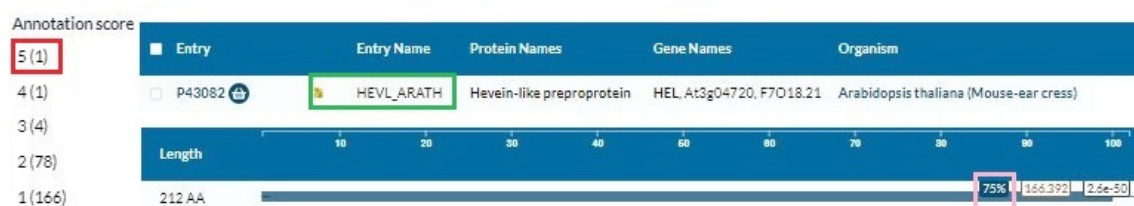


Рис. 2. Результат работы алгоритма BLAST (красный – ранг аннотаций, зеленый – имя протеина в базе данных, розовый – процент сходства)

Fig. 2. Result of the BLAST algorithm (red – the rank of annotations, green – the name of the protein in the database, pink – the percentage of similarity)

³National Center for Biotechnology Information [Электронный ресурс]. URL: <https://www.ncbi.nlm.nih.gov/search/all/?term=WIN> (02.12.2022).

⁴Web site for bioinformatics annotation for plants [Электронный ресурс]. URL: <https://plants.ensembl.org/info/about/index.html> (02.12.2022).

⁵Database of primary protein sequences [Электронный ресурс]. URL: <https://www.uniprot.org> – database of primary protein sequences (02.12.2022).

⁶Database for annotations of proteins by EMBL's EBI // European Molecular Biology Laboratory European Bioinformatics Institute [Электронный ресурс]. URL: <https://www.ebi.ac.uk/> (02.12.2022).

⁷The Protein Information Resource by bioinformatics [Электронный ресурс]. URL: <https://proteininformationresource.org/> (02.12.2022).



Рис. 3. Результаты выравнивания HEVL с Wound-induced Protein (фиолетовым цветом выделены консервативные участки, зеленым – предсказанные аминокислоты; цифрами обозначен номер последовательности)

Fig. 3. Results of alignment of HEVL with Wound-induced protein (conserved regions are highlighted in purple, predicted amino acids are highlighted in green; numbers indicate the sequence number)

не учитывались. Однако в зависимости от поставленных исследователями иных задач при применении данной методологии могут быть использованы и другие метрики.

По результату выравнивания выяснилось, что WIN имеет большое количество участков, консервативных по отношению к HEVL (рис. 3). Также была составлена матрица идентичности, равная 73,53% (рис. 4). Исходя из этого, делается вывод о высоком уровне гомологии, что в таком случае позволяет экстраполировать уже детектированные SNPs с HEVL на соевый протеин WIN.

В дальнейшем среди общего количества известных однонуклеотидных полиморфизмов для *Arabidopsis thaliana* был проведен отбор по совпа-

дениям, выявленным в результате выравнивания (таблица).

В итоге среди общего количества консервативных участков было отобрано 5 аминокислот в кодонах, в которых возможны однонуклеотидные полиморфизмы, имеющие следующие порядковые номера: 11, 64, 73, 76, 86 у фермента сои WIN.

ВЫВОДЫ

В ходе проведенного исследования по нашей методологии среди последовательности аминокислот было найдено 5 аминокислотных остатков, входящих в консервативные участки и вызванных однонуклеотидными полиморфизмами в соответствующем гене. Эти данные могут быть использованы в селекционном процессе для последующего получения нового сорта растений с устойчивостью к фитопатогенным грибкам. Однако следует учесть, что предсказание локализации однонуклеотидных полиморфизмов *in silico* носит теоретический характер и для дальнейшего подтверждения потребуется применение технически сложного оборудования и дорогостоящих методов. В лабораториях для детекции SNPs требуется тщательный мониторинг большого количества образцов разных сортов. Описанная в статье методика необходима для проведения искусственной индукции предсказанных мутаций. Дальнейшим этапом будет являться выявление изменений молекулярных свойств с использованием методов рентгеноструктурного анализа или криоэлектронной микроскопии. Также стоит учитывать, что ручной поиск полиморфизмов в предложенном методе может быть успешным, если первичные последовательности аминокислот короткие (имеют длину не более 250 букв), в случае работы с более длинными строками рекомендуется автоматизация данного процесса. Разработанная методика позволит сократить трудозатраты на нахождение полиморфизмов, а также нужное на это время благодаря тому, что отпадает необходимость полного исследования генома *in vitro*.

Percent Identity Matrix



Рис. 4. Матрица идентичности HEVL и WIN, полученная в результате выравнивания

Fig. 4. Identity matrix HEVL and WIN resulting from alignment

Локализация вариативных аминокислот
Localization of variable amino acids

Порядковый номер АМК в HEVL	Порядковый номер АМК в Wound-induced Protein	Заменяемые аминокислоты	Кодон (полуужирным помечен SNPs)
117	11	W-C	TGG/TG I
170	64	I-V	ATA/G T A
179	73	Q-K	CAA/ A AA
182	76	L-I	CTC/ A TC
192	86	G-D	GGC/G A C

REFERENCES

1. Lander E. S. Initial sequencing and analysis of the human genome. *Nature*. 2001;409:860-921. <https://doi.org/10.1038/35057062>.
2. Makki R. M., Saeedi A. A., Khan T. K., Ali H. M., Ramadan A. M. Single nucleotide polymorphism analysis in plastomes of eight *Catharanthus roseus* cultivars. *Biotechnology & Biotechnological Equipment*. 2019;33(1):419-428. <https://doi.org/10.1080/13102818.2019.1579671>.
3. Sukhumsirichart W. Polymorphisms. In: *Genetic Diversity and Disease Susceptibility*. Liu Ya. (ed.). 2018, 164 p. <https://doi.org/10.5772/intechopen.76728>.
4. Morgil H., Gercek Y. C., Tulum I. Single nucleotide polymorphisms (SNPs) in plant genetics and breeding. In: *The Recent Topics in Genetic Polymorphisms*. Çalışkan M., Erol O., Öz G. C. (eds.). 2020, 148 p. <https://doi.org/10.5772/intechopen.91886>.
5. Achard F., Butruille M., Madjarac S., Nelson P. T., Duesing J., Laffont J.-L., et al. Single nucleotide polymorphisms facilitate distinctness-uniformity-stability testing of soybean cultivars for plant variety protection. *Crop Science*. 2020;60(5):2280-2303. <https://doi.org/10.1002/csc2.20201>.
6. Freudenthal J. A., Ankenbrand M. J., Grimm D. G., Korte A. GWAS-flow: a GPU accelerated framework for efficient permutation based genome-wide association studies. *BioRxiv*. 2019;1:783100. <https://doi.org/10.1101/783100>.
7. Fan C., Zhai H., Wang H., Yue Y., Zhang M., Li J., et al. Identification of QTLs controlling grain protein concentration using a high-density SNP and SSR linkage map in barley (*Hordeum vulgare* L.). *BMC Plant Biology*. 2017;17:122. <https://doi.org/10.1186/s12870-017-1067-6>.
8. Zhang W., Xu W., Li S., Zhang H., Liu X., Cui X., et al. GmAOC₄ modulates seed germination by regulating JA biosynthesis in soybean. *Theoretical and Applied Genetics*. 2022;135(2):439-447. <https://doi.org/10.1007/s00122-021-03974-0>.
9. Mathur R., Rana B. S., Jha A. K. Single nucleotide polymorphism (SNP). In: *Encyclopedia of Animal Cognition and Behavior*. Vonk J., Shackelford T. (eds.). Cham., Springer; 2018, 7539 p. https://doi.org/10.1007/978-3-319-47829-6_2049-1.
10. Tchaikovskii V., Desnick R. J., Bishop D. F. Molecular expression, characterization and mechanism of ALAS₂ gain-of-function mutants. *Molecular Medicine*. 2019;24(25):4. <https://doi.org/10.1186/s10020-019-0070-9>.
11. Martynenko N., Lavrentieva S., Tarasova O. Influence of Septoria glycines Hemmi on the enzymes' activity of hydrolytic Glycines max (L.) Merr seed complex. *Fundamental and Applied Research in Biology and Agriculture: International Scientific and Practical Conference*. 2021;254:02020. <https://doi.org/10.1051/e3sconf/202125402020>.
12. Howe K. L., Contreras-Moreira B., De Silva N., Maslen G., Akanni W., Allen J., et al. Ensembl genomes 2020 – enabling non-vertebrate genomic research. *Nucleic Acids Research*. 2020;48(D1):D689-D695. <https://doi.org/10.1093/nar/gkz890>.
13. Rensink W. A., Buell C. R. Arabidopsis to rice. Applying knowledge from a weed to enhance our understanding of a crop species. *Plant Physiology*. 2004;135(2):622-629. <https://doi.org/10.1104/pp.104.040170>.
14. Coelho S. M., Peters A. F., Charrier B., Roze D., Destombe C., Valero M., et al. Complex life cycles of multicellular eukaryotes: new approaches based on the use of model organisms. *Gene*. 2007;406(1-2):152-170. <https://doi.org/10.1016/j.gene.2007.07.025>.
15. Meyerowitz E. M. Prehistory and history of Arabidopsis research. *Plant Physiology*. 2001;125(1):15-19.
16. Leonelli S. Arabidopsis, the botanical Drosophila: from mouse cress to model organism. *Endeavour*. 2007;31(1):34-38. <https://doi.org/10.1016/j.endeavour.2007.01.003>.
17. Lan Y., Sun R., Ouyang J., Ding W., Kim M. J., Wu J., et al. AtMAD: Arabidopsis thaliana multi-omics association database. *Nucleic Acids Research*. 2021;49(D1):D1445-D1451. <https://doi.org/10.1093/nar/gkaa1042>.
18. Jampala P., Garhewal A., Lodha M. Functions of long non-coding RNA in Arabidopsis thaliana. *Plant Signaling & Behavior*. 2021;16(9):1925440. <https://doi.org/10.1080/15592324.2021.1925440>.
19. Yu X., Vandivier L. E., Gregory B. D. NAD-seq for profiling the NAD⁺ capped transcriptome of Arabidopsis thaliana. *STAR Protocols*. 2021;2(4):100901. <https://doi.org/10.1016/j.xpro.2021.100901>.

ИНФОРМАЦИЯ ОБ АВТОРАХ

П. Д. Тимкин,
 младший научный сотрудник,
 Всероссийский научно-исследовательский
 институт сои,
 675000, г. Благовещенск, Игнатьевское шоссе,
 19, Российская Федерация,
 tpd@vniisoi.ru
<https://orcid.org/0000-0001-6655-1049>

INFORMATION ABOUT THE AUTHORS

Pavel D. Timkin,
 Junior Researcher,
 All-Russian Research Institute of Soybean,
 19, Ignatievskoe Shosse, Blagoveshchensk,
 675000, Russian Federation,
 tpd@vniisoi.ru
<https://orcid.org/0000-0001-6655-1049>

А. А. Пензин,
младший научный сотрудник,
Всероссийский научно-исследовательский
институт сои,
675000, г. Благовещенск, Игнатьевское шоссе,
19, Российская Федерация,
paa@vniiso.ru
<https://orcid.org/0000-0002-8578-9818>

Вклад авторов

Все авторы сделали эквивалентный вклад
в подготовку публикации.

Конфликт интересов

Авторы заявляют об отсутствии конфликта
интересов.

*Все авторы прочитали и одобрили
окончательный вариант рукописи.*

Информация о статье

Поступила в редакцию 16.11.2022.
Одобрена после рецензирования 29.11.2022.
Принята к публикации 30.11.2022.

Andrey A. Penzin,
Junior Researcher,
All-Russian Research Institute of soybean,
19, Ignatievskoe Shosse, Blagoveshchensk,
675000, Russian Federation,
paa@vniiso.ru
<https://orcid.org/0000-0002-8578-9818>

Contribution of the authors

The authors contributed equally to this article.

Conflict interests

The authors declare no conflict of interests regarding
the publication of this article.

*The final manuscript has been read and approved by
all the co-authors.*

Information about the article

The article was submitted 16.11.2022.
Approved after reviewing 29.11.2022.
Accepted for publication 30.11.2022.