

НОВОЖИЛОВ Д. О.<sup>✉</sup>, ГОРДИНСЬКИЙ С. О., ПОСТОВОЙТОВА А. С., РАБОКОНЬ А. М., ПІРКО Я. В., БЛЮМ Я. Б.

ДУ «Інститут харчової біотехнології та геноміки НАН України»,  
Україна, 04123, м. Київ, вул. Байди-Вишневецького, 2а, ORCID: 0009-0008-1491-7166, 0009-0005-7369-3359, 0000-0003-3768-5763, 0000-0002-6249-1824, 0000-0003-1887-5406, 0000-0001-7078-7548, 0000-0001-7078-7548

<sup>✉</sup> [novozhylovd@gmail.com](mailto:novozhylovd@gmail.com)

## АЛГОРИТМ СТВОРЕННЯ ІЛР-МАРКЕРІВ ДЛЯ МОЛЕКУЛЯРНО-ГЕНЕТИЧНОГО АНАЛІЗУ *TRITICUM DURUM* ТА ІНШИХ ВИДІВ ЗЛАКІВ

**Мета.** Розроблення алгоритму створення ІЛР-маркерів, придатних для селекційних досліджень пшениці твердої (*Triticum durum*) та деяких інших видів злаків. **Методи.** Використання засобів класичної біоінформатики для пошуку сайтів розташування потенційних інтронів цільових генів у *T. durum*, придатних для використання у якості ІЛР-маркерів на основі гомології контигів *T. durum* і кодуєчих послідовностей рису (*Oryza sativa*). Аналіз екзон-інтронної структури гомологічних генів у *Oryza sativa*, *Hordeum vulgare*, *Aegilops tauschii* та *Triticum aestivum* з наступним підбором відповідних праймерів до потенційних ІЛР-маркерів. **Результати.** Визначено ряд потенційних сайтів розташування цільових інтронів у послідовностях контигів *T. durum*. До них підібрані вироджені праймери із врахуванням аналізу екзон-інтронної структури відповідних гомологічних генів у *O. sativa*, *H. vulgare*, *A. tauschii* та *T. aestivum*. **Висновки.** Розроблено алгоритм створення ІЛР-маркерів, який може бути використаний у молекулярно-генетичних дослідженнях рослин у випадку відсутності або неповного сиквенованого геному досліджуваного виду і браку інформації про екзон-інтронну структуру їх генів. Визначено ряд потенційних ІЛР-маркерів, які можуть бути використані для злаків, зокрема *T. durum*.

**Ключові слова:** генетичні маркери, ІЛР-маркери, *Triticum durum*, CDS, екзон-інтронна структура генів, вироджені праймери.

ДНК-маркери є доступним та корисним інструментом для маркування господарсько цінних ознак у рослин, зокрема злаків. Впровадження методів молекулярного маркування значно пришвидшує селекцію ліній в процесі створення нових сортів. Однак актуальним залишається питання щодо пошуку нових ген-

специфічних ДНК-маркерів, які б були інформативними та доступними для генотипування. Багатообіцяючим напрямком розроблення генспецифічних ДНК-маркерів є метод оцінки поліморфізму довжини інтронів (Intron Length Polymorphism, ILP) [1–3].

Для систематизації та полегшення пошуку розроблених ІЛР-маркерів була створена онлайн база даних PIP (Potential Intron Polymorphism), в якій розміщувалась детальна інформація щодо PIP-маркерів та їх взаємозв'язку у різних видів рослин. PIP-маркери являють собою специфічні ІЛР-маркери, які створюються через дизайн праймерів до певних ділянок EST (expressed sequence tags). Пошук видоспецифічних PIP-маркерів раніше можна було здійснити через PIP базу даних, використавши посилання <http://ibi.zju.edu.cn/pgl/pip/> [4]. На жаль, сьогодні ця база даних вже не є функціонує, тому метою роботи було розроблення власного протоколу пошуку ІЛР-маркерів у геномах рослин та оцінка можливості його застосування в рамках генетико-селекційних досліджень *Triticum durum* та деяких інших видів злаків.

### Матеріали і методи

Геном та кодуєчі послідовності (CDS) *Oryza sativa Japonica Group* взяті з бази даних PlantGDB [5], а послідовності *T. durum* – з бази даних GenBank [6]. Вирівнювання CDS *O. sativa* проти геному *O. sativa* проводився з використанням інструменту Splign [7], а вирівнювання контигів *T. durum* та CDS *O. sativa* – з використанням MegaBLAST за допомогою програмного забезпечення NCBI Genome Workbench [8]. Фільтрування результатів і робота з таблицями проводилась у Microsoft Excel. Пошук гомологічних послідовностей до екзонних ділянок, що оточують цільові інтрони, здійснювали вбудованими засобами blastn бази даних GenBank.

© НОВОЖИЛОВ Д. О., ГОРДИНСЬКИЙ С. О., ПОСТОВОЙТОВА А. С., РАБОКОНЬ А. М., ПІРКО Я. В., БЛЮМ Я. Б.

Консенсусні послідовності були створені за допомогою інструменту Cons програмного пакету EMBOSS при параметрі множинності, що дорівнював кількості вихідних послідовностей [9]. Праймери підбирали за допомогою онлайн програми Primer3Plus за умов максимальної виродженості  $N - 2$  [10]. При підборі вироджених праймерів враховували рекомендації щодо їх розмірів та розташування дегенерованих залишків [11].

### Результати та обговорення

У цілому загальну схему створення ILP-маркерів представлено на рисунку. Вона передбачає залучення певних біоінформатичних підходів та використання доступних геномних баз даних. Ця схема була застосована нами для розроблення ILP-маркерів у *T. durum*.

Через брак даних щодо розташування інтронів у геномі *T. durum*, першочерговим завданням для пошуку потечійних ILP-маркерів, був вибір референсного геному. Як такий було обрано геном рису *O. sativa* – добре вивченої модельної злакової рослини з великою кількістю анотованих генів. Базуючись на гомології були визначені позиції розташування інтронів у референтному геномі *O. sativa* та здійснено пошук відповідних їм позицій у послідовностях контигів *T. durum*. Для цього проведено вирівнювання послідовностей CDS *O. sativa* до його геному за допомогою алгоритмів Splign, що дозволило визначити межі екзонів.

Наступним кроком було проведення вирівнювання послідовностей контигів *T. durum* до послідовностей CDS *O. sativa* засобами blastn. В якості гомологічних відібрано послідовності, ідентичність яких складала щонайменше 80 %, а перекриття не менше 200 пар основ (п. о.). Таким чином було визначено найближчі гомологи серед кодуєчих послідовностей *O. sativa* і контигів *T. durum* та співставлені аналогічні позиції з'єднання екзонів. Був проведений аналіз структури інтронів геномних ділянок *O. sativa*, що відповідають CDS, гомологічним до контигів *T. durum*. Визначено цільові інтрони, довжина яких становить від 200 до 400 п. о., що є оптимальним для створення ILP-маркерів.

Для подальшого підбору праймерів було відібрано послідовності екзонних ділянок, що оточують цільові інтрони. Довжина таких ділянок складає 200 п. о., по 100 п. о. з кожного боку від цільового інтрона. Засобами blastn прове-

дене вирівнювання проти бази даних GenBank для пошуку гомологічних послідовностей у *H. vulgare*, *A. tauschii* та *T. aestivum*. Проаналізована екзон-інтронна структура генів, що відповідають визначеним ділянкам і скорегована довжина для тих з них, де відстань між сусідніми інтронами є меншою, ніж 100 п. о. Також проведено вирівнювання послідовностей екзонних ділянок, що оточують цільові інтрони у *O. sativa*, *H. vulgare*, *A. tauschii*, *T. aestivum*, імовірні інтрони у *T. durum* та визначено консенсусні послідовності. До цих консенсусних послідовностей були підібрані вироджені праймери.

Як приклад можна навести методику розроблення двох ILP-маркерів для *T. durum*. На першому етапі була встановлена гомологія між контигом 238 *T. durum* та локусом Os01g47530.1 рису, на основі якої було визначено місце розташування інтрону цільового гена. На другому етапі – відібрані послідовності ділянок екзонів, що його фланкують у *T. durum* і будуть в подальшому використані для підбору праймерів. Пошуком blastn проти бази даних GenBank знайдено гомологічні послідовності у генах мітоген-активованої протеїнкінази 8 (MAPK8) *H. vulgare*, *A. tauschii*, *T. aestivum* та *O. sativa*. Проведено аналіз екзон-інтронної структури цих генів, за результатами якого права фланкуюча ділянка скорочена до 83 п. о. задля недопущення потрапляння сусідніх інтронів до області підбору праймерів. Створені спільні профілі екзонних ділянок, що фланкують інтрон цільового гена для *T. durum*, *H. vulgare*, *A. tauschii*, *T. aestivum* та *O. sativa*. До визначених консенсусних послідовностей були підібрані вироджені праймери. Оптимальна їх комбінація, згідно рекомендацій по дизайну – forward: TGGANGAAAATGGAGGGAAG; reverse: CTTTGGGGNACTTGGGAAG.

Наступним прикладом може бути методика підбору ILP-маркера до одного з інтронів гена, що кодує протеосому субодиницю альфа 3 (PSMA3, Proteasome subunit alpha 3). Зокрема встановлено гомологію між контигом 365 *T. durum* та локусом Os01g59600.1 рису, на основі якої було визначено місце розташування одного з інтронів. Далі було відібрано ділянки екзонів, які його фланкують у *T. durum* і будуть використані для підбору праймерів. Пошуком blastn проти бази даних GenBank знайдено гомологічні послідовності у генах PSMA3 *H. vulgare*, *A. tauschii*, *T. aestivum* та *O. sativa*.

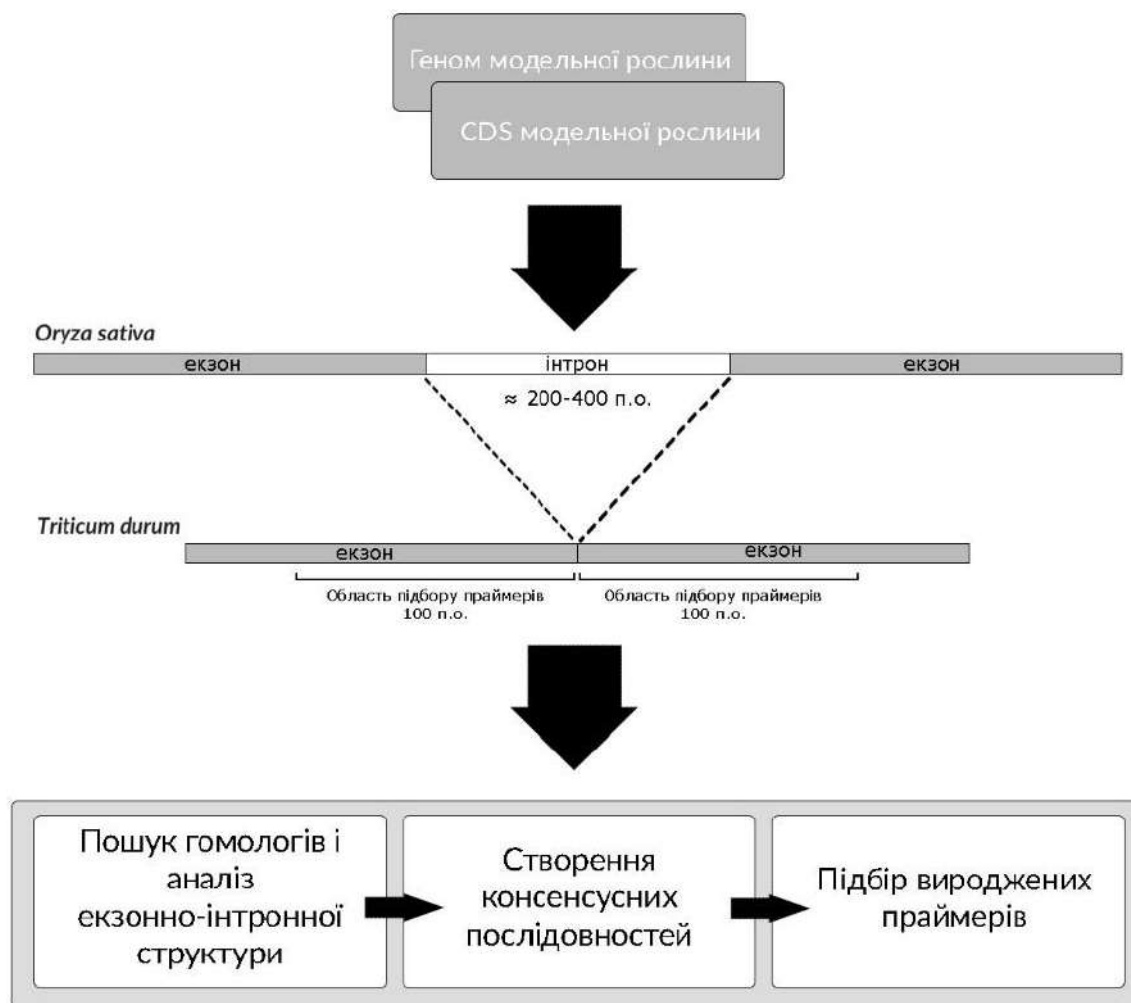


Рис. Схематичне зображення алгоритму створення ILP-маркерів.

Проведений аналіз екзон-інтронної структури цього гена, за результатами якого ліва фланкуюча ділянка скорочена до 84 п. о., а права – 48 п. о., задля недопущення потрапляння сусідніх інтронів у область підбору праймерів. Створені спільні профілі екзонних ділянок, що фланкують цільовий інтрон у *T. durum*, *H. vulgare*, *A. tauschii*, *T. aestivum* та *O. sativa*. До визначених консенсусних послідовностей підібрані вироджені праймери: forward – TGGCTGTGGAGTTATTCTTGG; reverse – TTTGCAGCNTGTCTNCCTTT.

Так само були визначені потенційні ILP-маркери до інтронів генів, які кодують: інгібітор цистеїн протеїнази 12 (Cysteine proteinase inhibitor 12), E3 убіхітин-протеїн лігазу (E3 ubiquitin-protein ligase, RGLG3), білок GIGANTEA, циклазу магній-протопорфірин IX монометилового

ефіру [окислювальну] (Magnesium-protoporphyrin IX monomethyl ester [oxidative] cyclase, CRD1, ZIP), тріозофосфат/фосфатний транслокатор (Triose phosphate/phosphate translocator, TPT), субодиницю протеасоми альфа 3 (Proteasome subunit alpha 3, PSMA3), каталітичну субодиницю 5 НАД<sup>+</sup>-ізоцитратдегідрогенази (Isocitrate dehydrogenase [NAD] catalytic subunit 5, IDH5), еноіл-КоА-редуктазу жирних кислот з дуже довгим ланцюгом (Very-long-chain enoyl-CoA reductase, ECR) (табл.).

Таким чином, розроблено алгоритм створення ILP-маркерів, який може бути використаний не тільки у молекулярно-генетичних дослідженнях *T. durum*, але й інших видів злаків, що створює підґрунтя до їх більш широкого використання у генетико-селекційних дослідженнях.

Таблиця. Результати пошуку потенційних ILP-маркерів та підібрані пари праймерів

Номер контигу <i>T. durum</i>	Гомологічний локус <i>O. sativa</i>	Номер інтрона	Білки, які кодуються відповідними генами <i>H. vulgare</i> , <i>A. tauschii</i> , <i>T. aestivum</i> та <i>O. sativa</i>	Оптимальна пара праймерів
77	Os01g16430.1	3	Інгібітор цистеїн протеїнази 12 (Cysteine proteinase inhibitor 12)	forward: NCCATGCTGTGAAATCAATCC reverse: NGCGAAGTCTTCAACNACCTC
191	Os01g73000.2	8	Е3 убіхітин-протеїн лігаза (E3 ubiquitin-protein ligase, RGLG3)	forward: TGGGNGATGGNCCATGG-GATGC reverse: GCTGCAAGNG-CAAATGCNGCCTC
238	Os01g47530.1	8	Мітоген-активована протеїнкіназа 8 (Mitogen-activated protein kinase 8, MAPK8)	forward: TGGANGAAAATGGAGGGAAG reverse: CTTTGGGGNACTTGGGAAG
240	Os01g08700.1	12	Білок GIGANTEA	forward: NCCAAGGCTTCANCTGCTAT reverse: TTCGCTGGTTGACTCTCCAC
246	Os01g17170.1	2	Циклаза магній-протопорфірин ІХ монометилового ефіру [окислювальна] Magnesium-protoporphyrin IX monomethyl ester [oxidative] cyclase (CRD1, ZIP)	forward: GTGGCTGAGATCTTCTCGCT reverse: TGGTCAA-GAANCCGAGGTCCA
246	Os01g17170.1	4	Циклаза магній-протопорфірин ІХ монометилового ефіру [окислювальна] Magnesium-protoporphyrin IX monomethyl ester [oxidative] cyclase (CRD1, ZIP)	forward: NACACCAAA-GAATTCGACATG reverse: NACCATCCTGTCTANCTTCCTC
299	Os01g13770.1	4	Тріозофосфат/фосфат-ний транслокатор (Triose phosphate/phosphate translocator, TPT)	forward: TGGAGCCCTTCTTCAATGCA reverse: TGCCATTGAAACACCAAGCA
365	Os01g59600.1	4	Субодинаця протеасоми альфа 3 (Proteasome subunit alpha 3, PSMA3)	forward: TGGCTGTGGAGTTATTCTTGG reverse: TTTGCAGCNTGTCTNCCTTT
655	Os01g16900.3	7	Каталітична субодинаця 5 НАД <sup>+</sup> -ізоцитрат-дегідрогенази (Isocitrate dehydrogenase [NAD] catalytic subunit 5, IDH5)	forward: GAAGGTGGCATTGTCTGGC reverse: TTGTGGATCCGGTCTGCTTG
692	Os01g05670.1	1	Єноіл-КоА-редуктаза жирних кислот з дуже довгим ланцюгом (Very-long-chain enoyl-CoA reductase, ECR)	forward: NATGAAGGTNTCCGTCGTGT reverse: NGATGGCNTCCTGCAGATCCG

### Висновки

Розроблений алгоритм дозволяє створювати ILP-маркери, які можуть бути використані у молекулярно-генетичних дослідженнях у випадку відсутності або неповного секвенува-

ного геному певних видів рослин і браку інформації про екзон-інтронну структуру їх генів. Визначено ряд потенційних ILP-маркерів, що можуть бути використані для злків, включно з *T. durum*.

## References

1. Wang X., Zhao X., Zhu J., Wu W. Genome-wide investigation of intron length polymorphisms and their potential as molecular markers in rice (*Oryza sativa* L.). *DNA Res.* 2005. Vol. 12 (6). P. 417–427. doi: 10.1093/dnares/dsi019.
2. Choi H. K., Kim D., Uhm T., Limpens E., Lim H., Mun J. H., Kalo P., Penmetsa R. V., Seres A., Kulikova O., Roe B. A., Bisseling T., Kiss G. B., Cook D. R. A sequence-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *M. sativa*. *Genetics.* 2004. Vol. 166 (3). P. 1463–1502. doi: 10.1534/genetics.166.3.1463.
3. Rabokon A. N., Demkovich A. E., Pirko Ya. V., Blume Ya. B. Studing of  $\beta$ -tubulin gene intron length polymorphism of *Triticum aestivum* L. and *Hordeum vulgare* L. varieties. *Factors in Experimental Evolution of Organisms.* 2015. Vol. 17. P. 82–86. [in Ukrainian]
4. Yang L., Jin G., Zhao X., Zheng Y., Xu Z., Wu W. PIP: a database of potential intron polymorphism markers. *Bioinformatics.* 2007. Vol. 23 (16). P. 2174–2177. doi: 10.1093/bioinformatics/btm296.
5. Duvick J., Fu A., Muppirala U., Sabharwal M., Wilkerson M. D., Lawrence C. J., Lushbough C., Brendel V. PlantGDB: a resource for comparative plant genomics. *Nucl. Acids Res.* 2008. Vol. 36. P. 959–965. doi: 10.1093/nar/gkm1041.
6. Sayers E.W., Cavanaugh M., Clark K., Ostell J., Pruitt K.D., Karsch-Mizrachi I. GenBank. *Nucl/ Acids Res.* 2020. Vol. 48 (D1). P. 84–86. doi: 10.1093/nar/gkz956.
7. Kapustin Y., Souvorov A., Tatusova T., Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct.* 2008. Vol. 3. P. 3:20. doi: 10.1186/1745-6150-3-20.
8. Kuznetsov A., Bollin C.J. NCBI Genome Workbench: Desktop software for comparative genomics, visualization, and GenBank data submission. *Methods Mol. Biol.* 2021. Vol. 2231. P. 261–295. doi: 10.1007/978-1-0716-1036-7\_16.
9. Rice P., Longden I., Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 2000. Vol. 16 (6). P. 276–277. doi: 10.1016/s0168-9525(00)02024-2.
10. Untergasser A., Cutcutache I., Koressaar T., Ye J., Faircloth B. C., Remm M., Rozen S. G. Primer3--new capabilities and interfaces. *Nucl. Acids Res.* 2012. Vol. 40 (15). P. e115. doi: 10.1093/nar/gks596.
11. Iserte J. A., Stephan B. I., Goñi S. E., Borio C. S., Ghiringhelli P. D., Lozano M. E. Family-specific degenerate primer design: a tool to design consensus degenerated oligonucleotides. *Biotechnol. Res. Int.* 2013. 2013: 383646. doi: 10.1155/2013/383646.

**NOVOZHYLOV D. O., HORDYNSKYI S. O., POSTOVOITOVA A. S., RABOKON A. M., PIRKO Ya. V., BLUME Ya. B.**

*Institute of Food Biotechnology and Genomics NAS of Ukraine, Ukraine, 04123, Kyiv, Baidy-Vyshnevetzkoho str., 2A*

## ALGORITHM FOR CREATING ILP MARKERS FOR MOLECULAR GENETIC ANALYSIS OF TRITICUM DURUM AND OTHER TYPES OF CEREALS

**Aim.** Development of an algorithm for creating ILP markers suitable for breeding studies of durum wheat (*Triticum durum*) and some other types of cereals. **Methods.** Use of classical bioinformatics tools to locate potential intron sites of target genes in *T. durum* suitable for use as ILP markers based on homology of *T. durum* contigs and rice (*Oryza sativa*) coding sequences. Analysis of exon-intron structure of homologous genes in *O. sativa*, *Hordeum vulgare*, *Aegilops tauschii* and *Triticum aestivum* followed by selection of appropriate primers for potential ILP markers. **Results.** A number of potential location sites of target introns in *T. durum* contig sequences were identified. Degenerate primers were designed for them, taking into account the analysis of the exon-intron structure of the corresponding homologous genes in *O. sativa*, *H. vulgare*, *A. tauschii* and *T. aestivum*. **Conclusions.** An algorithm for the creation of ILP markers was developed, which can be used in molecular genetic studies of plants in the case of the absence or incomplete sequenced genome of the studied species and the lack of information about the exon-intron structure of their genes. A number of potential ILP markers that can be used for cereals, in particular *T. durum*, have been identified. **Keywords:** genetic markers, ILP markers, *Triticum durum*, CDS, exon-intron gene structure, degenerated primers.