УДК 575.174.015.3:004.9:633.111.1 https://doi.org/10.7124/FEEO.v39.1845

ДЕМКОВИЧ А. Є.✉
ПІРКО Я. В.

ДУ «Інститут харчової біотехнології та геноміки НАН України»,
Україна, 04123, м. Київ, вул. Байди-Вишневецького, 2а
✉ *demand.dn@gmail.com*

УНІВЕРСАЛЬНИЙ ВЕБ-ІНСТРУМЕНТ ДЛЯ АВТОМАТИЧНОГО ВИЯВЛЕННЯ ІНТРОННИХ ДІЛЯНОК ГЕНІВ У СКЛАДНИХ ГЕНОМАХ ДЛЯ ГЕНЕРУВАННЯ ILP-МАРКЕРІВ (НА ПРИКЛАДІ *TRITICUM AESTIVUM* L.)

Мета. Розробити зручний та універсальний веб-інструмент для автоматичного виявлення в геномних даних інтронних ділянок генів разом із прилеглими короткими фрагментами екзонів (50 п. о.) з метою їх подальшого використання для генерування ILP (Intron Length Polymorphism) маркерів. Інструмент працює безпосередньо з файлами GFF3 і FASTA та покликаний спростити аналіз складних геномів, зокрема таких, як у пшениці. ***Результати.*** Було розроблено систему, що працює у хмарі без окремого сервера. Вона швидко обробляє великі обсяги даних: наприклад, аналіз однієї хромосоми пшениці зайняв лише 118 секунд. Завдяки оптимізації пошуку система ефективно знаходить потрібні генетичні ділянки навіть у великих геномах. Тестування на 2B хромосомі *Triticum aestivum* виявило 15 664 інтронних локусів у 4 005 генах, забезпечуючи високу щільність – 3,91 маркера на ген. ***Висновки.*** Створена система значно прискорює пошук генетичних маркерів (ILP-маркерів) і дозволяє отримувати їх у великій кількості навіть для складних геномів. Крім того, результати одразу придатні для подальшого дизайну праймерів, що економить час і спрощує роботу дослідників.

Ключові слова: генетичні маркери, ILP-маркери, *Triticum aestivum*, CDS, екзон-інтронна структура генів.

Сучасна генетика та селекція потребують економічно доцільних і надійних ДНК-маркерів. Традиційні (SSR) та високопродуктивні (SNP) системи мають суттєві обмеження через видоспецифічність, високу вартість розробки або потребу в складному обладнанні (Novozhylov et al., 2025; Biswas et al., 2024; Yang et al., 2007; Liu et al., 2026). Високоефективною альтернативою є кодомінантні маркери поліморфізму довжини інтронів (ILP) (Shen et al., 2024). Вони базуються на еволюційній варіабельності інтронів та консервативності екзонів, що дозволяє створювати універсальні ПЛР-праймери для широкого кола споріднених таксонів. Тривалий час ключовою перепоною для масового пошуку ILP-маркерів була алгоритмічна складність процесу. Відома база PIP (2007) (Yang et al., 2007), яка залежала від EST-даних (охоплювала лише експресовані гени), обмежувалася лише рослинами і наразі не функціонує. Сьогодні дослідники змушені писати складні програми або створювати вузькоспеціалізовані локальні бази, що є критичною проблемою при аналізі гігантських поліплоїдних геномів, таких як пшениця (16 Гб). З огляду на це, метою роботи стало створення універсального масштабованого хмарного веб-інструменту для автоматичної обробки інтронних ділянок (із консервативними екзонними фланками для дизайну праймерів) у режимі реального часу безпосередньо з GFF3 та FASTA-файлів будь-якого еукаріотичного організму.

Матеріали і методи

Для реалізації інструменту було використано екосистему Amazon Web Services (AWS) на основі досвіду впровадження масштабованих інструментів для аналізу Big Data, описаний у нещодавніх оглядах (Kadri et al., 2022; Djaffardjy et al., 2023; Langmead and Nellore, 2018). Маршрутизація запитів через Amazon API Gateway та використання AWS Lambda забезпечує миттєвий відгук інтерфейсу. Однак, враховуючи обме-

© 2026 Демкович А. Є., Пірко Я. В. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

ження Lambda щодо тривалості процесів, обчислювальне ядро було реалізоване на базі AWS Fargate. Оскільки аналіз еукаріотичних геномів може тривати від кількох секунд до десятків хвилин, систему було побудовано на асинхронній моделі. Для узгодженості роботи використано інноваційний підхід «Storage-as-a-Queue» на базі AWS S3. Замість складних інструментів для передачі даних, стан завдання фіксується через фізичне переміщення метаданих між директоріями сховища. Це дозволяє реалізувати надійну систему статусів, яку користувач бачить на своїй сторінці статистики. Доступ до інструменту реалізований як статичний файловий хостинг: http://introne-search.s3-website-us-east-1.amazonaws.com/ З метою оптимізації використання хмарних ресурсів імплементовано Fail-Fast валідацію. На відміну від старих підходів, які перевіряють дані лише після повного завантаження файлу, наш метод дозволяє аналізувати файл частинами – починаючи вже з його заголовка. Це означає, що система працює швидше й не витрачає зайві ресурси, що відповідає сучасним підходам до обробки геномних даних.

Обробка форматів GFF3 та FASTA традиційно вважається ресурсномісткою. Хоча інструменти типу bedtools (Quinlan and Hall, 2010) залишаються стандартом, останні дослідження (Djaffardjy et al., 2023; Langmead and Nellore, 2018) вказують на переваги вбудованих реляційних баз даних для швидкого пошуку інтервалів у хмарних контейнерах. Замість ітеративних циклів Python ($O(N^2)$) використано SQLite. Трансформація анотації в індексовані SQL-таблиці дозволяє виконувати пошук інтронів через операцію INNER JOIN таблиці екзонів самої з собою. Умовою злиття є ідентичність gene_id та суміжність номерів екзонів ($ID < sub > n < sub >$ та $ID < sub > n + 1 < sub >$). Такий підхід знижує обчислювальну складність до $O(N \log N)$, що є критичним для аналізу великих геномів еукаріот.

Для забезпечення інтеграції з інструментами дизайну праймерів (зокрема Primer3), розроблений інструмент автоматично додає до кожного інтрона по 50 п.о. консервативних екзонних ділянок. Це вирішує проблему «ізольованих інтронів», характерну для ранніх баз даних типу PIP. Згенеровані CSV-результати зберігаються в AWS S3 з активною політикою Lifecycle Management. Відповідно до стандартів безпеки даних та оптимізації витрат, файли мають 14-денний TTL (Time-to-Live), після чого автоматично видаляються.

Результати та обговорення

Для перевірки обчислювальної стійкості розробленого інструменту було обрано хромосому 2B *Triticum aestivum*. Вибір гексаплоїдної пшениці як тестової моделі зумовлений відносною складністю її геному (загальний обсяг понад 16 Гб), що характеризується високим рівнем плоїдності, насиченістю ретротранспозонами та численними паралогічними копіями генів. Для традиційних біоінформатичних інструментів, що базуються на послідовній обробці великих текстових файлів (GFF3 та FASTA), обробка таких масивів зазвичай супроводжується експоненціальним зростанням споживання оперативної пам'яті (RAM) та значними часовими витратами. Повний цикл обробки – від валідації віддаленого заголовка файлу до формування фінального масиву з 15 664 інтронних послідовностей – зайняв лише 118 секунд (табл. 1). Це підтверджує гіпотезу про те, що безсерверна модель є більш ефективною для епізодичних високонавантажених обчислень, ніж утримання дорогої стаціонарної інфраструктури або використання стандартних скриптових рішень.

Таблиця 1. Характестики обчислювальної ефективності інструменту (на прикладі *T. aestivum*, Chr 2B)

Етап обробки	**Технологія / Метод**	**Час (сек)**	**Метрика результату**
Валідація вхідних URL	AWS Lambda / Fail-Fast	2.4	Перевірка цілісності структур
Обробка даних із подальшою індексацією	SQLite / Fargate	42.1	4005 анотованих локусів
Ідентифікація інтронів	SQL INNER JOIN	35.8	15664 унікальні локуси
Виділення цільових ділянок із фланкуючими послідовностями	Python Range Requests	37.7	FASTA + 50 bp екзонні фланки
ЗАГАЛЬНИЙ ЧАС	SaaS Pipeline	118.0	Готовність до дизайну маркерів

Ідентифіковано 15 664 локуси у 4 005 генах, що демонструє середню частоту 3,91 інтрони на один ген. В умовах селекції пшениці, де гени часто мають ідентичні екзонні послідовності, саме інтрони стають основним джерелом Indel-поліморфізму для розробки специфічних маркерів. Висока щільність (до 4 потенційні маркери на локус) створює так звану «архітектурну надлишковість». Якщо один інтрон у цільовому гені (наприклад, гени стійкості до патогенів) виявляється консервативним у досліджуваних сортів, дослідник автоматично отримує доступ до послідовностей ще трьох локусів у межах того самого гена. Це радикально скорочує час на ітеративний пошук маркерів, що було головним обмеженням при використанні застарілих статичних баз даних типу PIP, де для більшості локусів був доступний лише один, часто фрагментарний, інтрон.

Для верифікації біологічної значущості ідентифікованих локусів було проведено функціональну анотацію 4005 асоційованих генів у системі PANTHER (Mi et al., 2021). Класифікація за молекулярною функцією виявила виражене домінування двох стратегічно важливих категорій: каталітичної активності (GO:0003824) – 883 гени (22,0%) та функцій зв'язування (GO:0005488) – 417 генів (10,4%).

Аналіз розподілу ідентифікованих генів за класами білків, які вони кодують, надав можливість деталізувати біохімічний потенціал розроблених маркерів. Найбільшу функціональну групу складали ферменти (PC:00262) – 758 генів (18,9%). Додатково було ідентифіковано значні кластери білкових модифікаторів (247 генів, 6,2%) та транспортерів (180 генів, 4,5%), що вказує на здатність інструменту маркувати складні системи посттрансляційної регуляції та мембранного транспорту, які є критичними для адаптації рослин до абіотичних стресів, таких як посуха чи засолення.

Висока частка некласифікованих локусів у всіх типах аналізу PANTHER (від 50,5% у Protein Class до 97,5% у Pathways) привертає особливу увагу. У класичних дослідженнях, орієнтованих на відомі метаболічні шляхи, ці гени часто ігнорувалися. Проте, з точки зору сучасної геноміки, цей масив представляє так звану «темну матерію» геному – гени, функція яких ще не встановлена, але які можуть бути детермінантами адаптивності до абіотичних змін або специфічної якості зерна.

Можливість ідентифікації інтронів у цих маловивчених ділянках перетворює запропонований веб-сервіс на важливий інструмент для генетичних досліджень. Це дає змогу дослідникам використовувати розроблені ILP-маркери як орієнтири для GWAS-аналізу та картування нових QTL у ділянках хромосом, які раніше залишалися поза межами аналізу через відсутність функціональної анотації в статичних базах даних.

Однією з головних проблем сучасних веб-інструментів є «мовчазна відмова» при завантаженні некоректних геномних збірок. У розробленому інструменті реалізовано модуль Fail-Fast валідації, інтегрований у AWS Lambda. На відміну від традиційних інструментів, які починають обробку і видають помилку лише через кілька хвилин після зчитування всього файлу, запропонований підхід базується на частковому завантаженні метаданих. Система перевіряє відповідність заголовків GFF3 та FASTA, наявність індексованих координат та цілісність посилань протягом перших 2.4 секунди (табл. 1). Це дозволяє уникнути зайвих витрат хмарних ресурсів і забезпечує миттєвий зворотний зв'язок для користувача. В умовах роботи з Big Data пшениці, де помилка в одному хромосомному файлі може призвести до збою всього інструменту, такий архітектурний фільтр є критично важливим для забезпечення стабільності сервісу. За умови збою реєстри подій Fargate аавтоматично прикріплюються до карточки аналізу і є доступними для перегляду а данні запиту до виправлення. Порівняння розробленого веб-сервісу із найбільш розповсюдженими інструментами ідентифікації ILP-маркерів наведено в (табл. 2).

У рамках асинхронної моделі координації, реалізованої через Amazon API Gateway та черги Amazon S3, кожному завданню автоматично присвоюється унікальний ідентифікатор (UUID), що забезпечує надійне відстеження та кероване виконання обробки даних. Це дозволяє користувачу закрити браузер і повернутися до результатів пізніше. Важливою інновацією є впровадження політики життєвого циклу даних (TTL): результати зберігаються лише 14 днів, після чого автоматично видаляються. Це відповідає сучасним стандартам безпеки геномних даних та оптимізації витрат на хмарне сховище, що вигідно відрізняє цей сервіс від статичних баз, які часто містять застарілі версії збірок.

Таблиця 2. Порівняльна характеристика систем розробки ILP-маркерів

Критерій порівняння	База даних PIP (2007)	CLI інструменти (bedtools)	Розроблений веб-сервіс
Джерело даних	Статичні EST-бібліотеки	Локальні GFF3/FASTA	Динамічні URL-посилання
Обробка Big Data	Неможлива (ліміт за обсягом)	Залежить від RAM станції	Еластична (AWS Serverless)
Алгоритм пошуку	BLAST alignment (gaps)	Аналіз текстових файлів	Реляційна алгебра (SQLite)
Вихідний формат	Лише інтронна послідовність	Координати / FASTA	CSV + 50 bp екзонні ділянки
Валідація даних	Відсутня	Ручна (користувачем)	Fail-Fast (API Gateway)

На відміну від класичних CLI-інструментів (наприклад, bedtools), де дослідник отримує лише нуклеотидну послідовність інтрона, розроблений алгоритм автоматично розширює координати на 50 нуклеотидів у кожну сторону від точок сплайсингу. Це гарантує наявність стабільних зон для посадки праймерів у консервативних екзонних ділянках, що критично для успішної ампліфікації інтронних ділянок у геномах із високою варіабельністю. Отриманий CSV-файл, що містить 50 п.о. консервативних екзонних ділянок, є готовою матрицею для прямого імпорту в Primer3 або Primer-BLAST без додаткової біоінформатичної обробки (Gheno et al., 2026).

Висновки

1. Розроблено та імплементовано безсерверну систему на базі AWS, здатну обробляти складні повногеномні анотації (на прикладі пшениці, 16 Гб), що забезпечує високу швидкість аналізу незалежно від локальних обчислювальних ресурсів користувача.

2. На прикладі хромосоми 2B *T. aestivum* доведено здатність алгоритму генерувати надлишкові бібліотеки маркерів із середньою щільністю 3,91 локуси на ген, що вказує на високу роздільну здатність інструменту та гарантує надійність розробки специфічних Indel-маркерів для складних поліплоїдних об’єктів.

3. Мультимодальна анотація верифікувала повне покриття геному: від мембранних рецепторів стійкості (31,9%) до фундаментальних шляхів метаболізму (реплікація ДНК, убіквітиновий шлях та ін.), що підтверджує біологічну валідацію та репрезентативність отриманих даних.

4. Інтегрована видача послідовностей разом із консервативними ділянками екзонів (50 п. о.) усуває «біоінформатичний бар’єр», забезпечуючи прямий перехід до дизайну праймерів та радикальне прискорення етапу ПЛР-діагностики в лабораторних умовах.

Інформація про фінансову підтримку, подяки: відсутня.

Конфлікт інтересів: автори заявляють про відсутність конфлікту інтересів.

References

1. Aleksander S. A., Balhoff J. P., Carbon S., Cherry J. M., Ebert D., Feuermann M., Gaudet P., Harris N. L., Hill D. P., Kalita P., Lee R., Mi H., Moxon S., Mungall C. J., Muruganujan A., Mushayahama T., Sternberg P. W., Thomas P. D., van Auken, K., … Westerfield M. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Research*. 2026. Vol. 54 (D1). P. D1779–D1792. https://doi.org/10.1093/nar/gkaf1292.
2. Biswas M. K., Biswas D., Yi G., Deng G. The Musa Marker Database: A Comprehensive Genomic Resource for the Improvement of the Musaceae Family. *Agronomy*. 2024. Vol. 14 (4). P. 838. https://doi.org/10.3390/agronomy14040838.
3. Djaffardjy M., Marchment G., Sebe C., Blanchet R., Belhajjame K., Gaignard A., Lemoine F., Cohen-Boulakia S. Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems. *Computational and Structural Biotechnology Journal*. 2023. Vol. 21. P. 2075–2085. https://doi.org/10.1016/j.csbj.2023.03.003.
4. Gheno B. P., Fröder H., Lopes T. S., Rissi V. B., de Avila e Silva S. Steps for primer development using open-access platforms. *Biologia*. 2026. Vol. 81 (1). Article number 24. https://doi.org/10.1007/s11756-025-02088-y.
5. Kadri S., Sboner A., Sigaras A., Roy S. Containers in Bioinformatics. *The Journal of Molecular Diagnostics*. 2022. Vol. 24 (5). P. 442–454. https://doi.org/10.1016/j.jmoldx.2022.01.006.
6. Langmead B., Nellore A. Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*. 2018. Vol. 19 (4). P. 208–219. https://doi.org/10.1038/nrg.2017.113.

7. Liu S., Geng S., Dreisigacker S. Genomics in wheat improvement: Progress and perspectives. *The Plant Genome*. 2026. Vol. 19 (1). https://doi.org/10.1002/tpg2.70176.
8. Mi H., Ebert D., Muruganujan A., Mills C., Albou L.-P., Mushayamaha T., Thomas P. D. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*. 2021. Vol. 49 (D1). P. D394–D403. https://doi.org/10.1093/nar/gkaa1106.
9. Novozhylov D. O., Rabokon A. M., Blume R. Y., Demkovych A. Y., Pryvalikhin S. M., Mykhailova D. M., Pirko Ya. V., Blume Ya. B. Design of ILP markers for β-tubulin genes analysis in basidiomycete fungi. *Faktori Eksperimental'noi Evolucii Organizmiv*. 2025. Vol. 36. P. 75–81. https://doi.org/10.7124/FEEO.v36.1716.
10. Quinlan A. R., Hall I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010. Vol. 26 (6). P. 841–842. https://doi.org/10.1093/bioinformatics/btq033.
11. Shen Y., He X., Zu F., Huang X., Yin S., Wang L., Geng F., Cheng X. Development of Genome-Wide Intron Length Polymorphism (ILP) Markers in Tea Plant (*Camellia sinensis*) and Related Applications for Genetics Research. *International Journal of Molecular Sciences*. 2024. Vol. 25 (6). P. 3241. https://doi.org/10.3390/ijms25063241.
12. 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. https://doi.org/10.1093/bioinformatics/btm296.

DEMKOVYCH A. Ye, PIRKO Ya. V.
Institute of Food Biotechnology and Genomics NAS of Ukraine,
Ukraine, 04123, Kyiv, Baidy-Vyshnevetskoho str., 2A

A VERSATILE WEB TOOL FOR THE AUTOMATIC IDENTIFICATION OF INTRONIC REGIONS OF GENES IN COMPLEX GENOMES FOR GENERATING ILP MARKERS (*TRITICUM AESTIVUM* L. EXAMPLE)

Aim. Develop a user-friendly and versatile web tool for the automatic identification of intronic regions of genes along with adjacent short exon fragments (50 bp) in genomic data for their subsequent use in generating ILP (Intron Length Polymorphism) markers. The tool works directly with GFF3 and FASTA files and is designed to simplify the analysis of complex genomes, particularly those of wheat. ***Results.*** A serverless cloud system was developed, capable of rapidly processing large volumes of data. For example, analyzing a single wheat chromosome took only 118 seconds. Thanks to optimized search algorithms, the system efficiently identifies target genetic regions even in large genomes. Testing on the 2B chromosome of *Triticum aestivum* revealed 15,664 intron loci across 4,005 genes, achieving a high density of 3.91 markers per gene. ***Conclusions.*** The developed system significantly accelerates the identification of genetic markers (ILP markers) and allows generation of large marker libraries even for complex genomes. Moreover, the results are immediately suitable for subsequent primer design, which saves time and simplifies researchers' work.
Keywords: genetic markers, ILP markers, *Triticum aestivum*, CDS, exon-intron gene structure.

Надходження статті: 5 травня 2026 р.
Прийняття до друку після рецензування: 15 червня 2026 р.
Публікація: 30 листопада 2026 р.