DOI: https://doi.org/10.18524/2307-4663.2025.3(65).345238

УДК 578.834

К. С. Онищенко[1], О. Ю. Зінченко[2], Б. С. Жуков[3]

[1]Вінницький національний медичний університет ім. М. І. Пирогова,
вул. Пирогова, 56, м. Вінниця, Вінницька область, 21018,
e-mail: cat.medichka@gmail.com
[2]Одеський національний університет імені І. І. Мечникова,
вул. Змієнка Всеволода, 2, м. Одеса, Одеська область, 65082,
e-mail: ozinchenko@onu.edu.ua
[3]Одеський національний технологічний університет,
вул. Канатна, 112, м. Одеса, Одеська область, 65039,
e-mail: brs.alchemist@gmail.com

РЕТРОСПЕКТИВНЕ МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ СВІТОВОГО МОНІТОРИНГУ ДИНАМІКИ ЛІНІЙ *BETACORONAVIRUS PANDEMICUM* ЯК ІНСТРУМЕНТ НАГЛЯДУ ЗА НОВІТНІМИ МЕДИЧНИМИ ЗАГРОЗАМИ

Мета. *Оцінити параметри профілів ліній SARS-CoV-2 (Betacoronavirus pandemicum) та встановити залежності між обсягами секвенування, часовими характеристиками моніторингу та різноманіттям ліній.* ***Методи.*** *Проаналізовано дані 11,33 млн. сіквенсів та 2703 ліній B. pandemicum із 176 країн за номенклатурою Pango. Визначали індекси концентрованості (HHI), різноманіття (H, HMM), рівномірності (J), тривалість спостереження (D) та інтенсивність секвенування (I). Застосовані методи кореляційного і регресійного аналізу із використанням відповідних бібліотек python.* ***Результати.*** *Виявлено сильні зв'язки між тривалістю та інтенсивністю моніторингу й числом ідентифікованих ліній. Визначені порогові значення для досягнення охоплення 90%, 95% і 99% різноманіття становили відповідно ~239, 321 і 411 сіквенсів. Після ~400–500 сіквенсів спостерігалося зниження ефективності, що відображало мінімальний приріст нових ліній. Маржинальна віддача секвенування знижувалася до <1 нової лінії на 1000 додаткових сіквенсів.* ***Висновки.*** *Часові параметри та інтенсивність секвенування визначають ефективність генетичного нагляду, тоді як рівномірність профілю залежить переважно від дизайну відбору зразків. Запропонована модель дозволяє встановлювати пороги секвенування та оптимізувати моніторинг.*

Ключові слова: COVID-19, Betacoronavirus pandemicum, лінії, дизайн-модель, секвенування, різноманіття, моніторинг.

Поява *Betacoronavirus pandemicum* та подальша пандемія COVID-19 (COrona VIrus Disease 2019) виявили масштабну вразливість глобальних систем охорони здоров'я та економіки до новітніх вірусних загроз. Окрім значного епідемічного тиску, пандемія продемонструвала виняткову здатність вірусу

швидко формувати генетичне різноманіття й специфічні лінії, що відрізнялися біологічними властивостями, епідемічною динамікою, вірулентністю та чутливістю до імунної відповіді [7, 8]. Значна генетична мінливість, що є характерною для *Coronaviridae*, дозволяє представникам цієї родини випадково набувати нових конфігурацій поверхневих білків, надаючи їм здатність ефективно зв'язуватися з клітинними рецепторами інших біологічних видів. Свого часу спалахи *SARS* та *MERS* [1, 14] наочно продемонстрували наслідки долання міжвидового бар'єру цією родиною вірусів, однак *B. pandemicum* переважив за впливом усі попередні варіанти. Така варіабельність зумовила нагальну потребу у створенні оперативних, стандартизованих і високочутливих інструментів молекулярно-генетичного нагляду, здатних забезпечувати раннє виявлення ключових варіантів та підтримувати прийняття управлінських рішень у сфері громадського здоров'я на глобальному рівні [6].

Важливим результатом глобальної мобілізації стало формування системи номенклатури Pango – єдиної та динамічної структури класифікації ліній *B. pandemicum*, яка інтегрує дані тисяч лабораторій у різних країнах та забезпечує уніфіковане відстеження еволюційних процесів вірусу в реальному часі. Використання Pango стало основою сучасного генетичного епіднагляду, яка підтримує ідентифікацію варіантів, оцінку ефективності протиепідемічних заходів, таргетований відбір зразків, коригування вакцинних стратегій і проведення філогенетичного аналізу на міжнародному рівні [10, 11].

Для систем охорони здоров'я існування цієї платформи забезпечує можливість розробки практичних рішень, що включають раннє попередження про потенційно особливо небезпечні VOC (Variant of Concern; варіанти, що викликають занепокоєння) та VOI (Variant of Interest; варіанти що викликають інтерес) варіанти, коригування алгоритмів тестування й секвенування, оцінювання ефективності вакцин і терапевтичних стратегій, підтримку таргетованих протиепідемічних заходів і планування ресурсів. Крім того, Pango створює прозорий канал комунікації між дослідниками та органами громадського здоров'я, тим самим підвищуючи швидкість та обґрунтованість протиепідемічних і профілактичних заходів на національному й міжнародному рівнях [2, 9, 11].

Попри значну кількість накопичених сіквенсів та широке використання інструментів генетичного моніторингу, сьогодні залишається недостатньо вивченим питання кількісної оцінки ефективності таких програм у просторі та часі. Поза увагою часто лишається визначення оптимального обсягу секвенувань, вплив часової структури на якість виявлення ліній, роль різних метрик різноманіття у формуванні репрезентативного профілю та пороги, після яких збільшення секвенувань перестає істотно підвищувати інформативність. Не менш важливою є проблема дизайн-ефекту, коли неоднорідність відбору зразків та кластеризація потоків суттєво змінюють ефективний розмір вибірки та точність оцінки переважання варіантів.

Ретроспективний аналіз світового моніторингу ліній *B. pandemicum* дає можливість сформувати математично обґрунтовану модель, яка описує ключові характеристики найбільших національних програм секвенування. Такий підхід дозволяє встановити порогові значення валового обсягу сіквенсів та

інтенсивності моніторингу, визначити динаміку кривої накопичення ліній та оцінити параметри, необхідні для надійного виявлення рідкісних варіантів. На основі отриманих метрик можливе формування узгодженого критерію якості, що здатний слугувати основою для проєктування майбутніх систем нагляду та оптимізації розподілу ресурсів, підвищуючи ймовірність раннього виявлення VOC та VOI.

Мета цієї роботи – здійснити ретроспективне математичне моделювання глобальної стратегії геномного моніторингу ліній *B. pandemicum* із використанням даних номенклатури Pango та визначити кількісні параметри ефективності, порогові значення секвенування, вплив часових компонент і дизайн-ефекту на якість профілю ліній. Особливу увагу приділено аналізу метрик концентрованості, різноманіття й рівномірності; сублінійній динаміці накопичення ліній; оцінці інтенсивності секвенування, необхідної для ідентифікації варіантів низької частоти; а також формуванню рекомендацій стосовно оптимізації національних і глобальних програм генетичного епіднагляду.

Представлена модель має потенціал для практичного застосування у сценарному плануванні майбутніх пандемічних загроз, особливо з боку вірусів родини *Coronaviridae*, та може бути адаптована до інших вірусних патогенів із високою швидкістю еволюції.

Матеріали і методи

Джерела даних. Для моделювання використано відкриту базу даних *pango-designation* (GitHub, https://github.com/cov-lineages/pango-designation), яка містить повний реєстр визначених ліній *B. pandemicum* та відповідні метадані ідентифікації. У дослідження включено дані за період від 24.12.2019 до 09.10.2024, актуальний стан яких фіксовано на 10.08.2025.

До аналізу увійшли 11 330 064 сіквенсів з 176 країн-учасниць моніторингу, 2703 епідеміологічні лінії *B. pandemicum*.

Відбір даних був тотальним, без виключення країн, незалежно від масштабів національних програм секвенування або їхньої регулярності.

Підготовка даних та формування профілів країн. Для кожної країни формували індивідуальний профіль ліній, який включав: 1) загальну кількість наданих сіквенсів; 2) кількість ідентифікованих ліній; 3) розподіл часток кожної лінії в межах національного набору; 4) часові межі першого та останнього зразка для кожної лінії; 5) часовий обсяг спостереження у межах країни.

Оскільки країни істотно відрізнялися за обсягом секвенувань, профілі були нормовані для забезпечення міжкраїнової порівнюваності.

Метрики профілю ліній. Для оцінки структури різноманіття застосовували три взаємодоповнювальні метрики:

1. Індекс Герфіндаля-Гіршмана (HHI)

Відображає концентрованість профілю та переважання окремих ліній:

$$HHI = \sum_{i=1}^{k} p_i^2,$$

де p_i – частка і-тої лінії.

2. Ентропія Шеннона (H) та коригована ентропія Міллера-Медоу (HMM)

Використовували для оцінки генетичного різноманіття. Для компенсації систематичного зсуву ентропії за умов нерівнозначних обсягів секвенувань і ідентифікованих ліній між країнами ентропію коригували поправкою Міллера-Медоу.

$$HMM = H + \frac{k-1}{2N},$$

де (k) – кількість ліній, (N) – кількість сіквенсів.

3. Рівномірність Пілоу (J)

Характеризує рівномірність розподілу часток ліній:

$$J = \frac{H}{\ln k}.$$

Підбір метрик здійснювали на основі аналізу джерел літератури [3, 4, 5, 12, 13].

Часові показники моніторингу. На основі часових міток сіквенсів визначали дві узагальнені часові характеристики:

1. Тривалість активного спостереження (D)

Різниця між датами першого та останнього отриманого зразка в країні.

2. Інтенсивність секвенування (I)

Середня кількість сіквенсів на добу у межах індивідуального календарного вікна:

$$I = \frac{N_{total}}{D}.$$

Ці метрики дозволяють оцінити безперервність та щільність генетичного моніторингу.

Моделі регресійної апроксимації. Для аналізу залежності кількості сіквенсів від параметрів профілю оцінювали 12 математичних моделей, серед яких представлено лінійну, логарифмічну, ступеневу, експоненційну, поліном 2-го та 3-го ступенів, обернену, кореневу, кубічну кореневу, квадратично лог-поліномну, Міхаеліса-Ментен та логіт-лог залежності. Моделі порівнювали за коефіцієнтом детермінації (R^2), стандартною похибкою апроксимації та адекватністю опису кривої накопичення.

Оцінка порогів виявлення різноманіття. Для кожної країни визначали мінімальний обсяг секвенувань (n), необхідний для досягнення заданої частки спостережуваного різноманіття (90%, 95%, 99%) на основі моделі:

$$E[L_i(n)] = \sum_{l:k_{iL}>0} (1 - (1 - p_{iL})^n)$$

де $E[S_i(n)]$ – очікуване число унікальних ліній у вибірці з n сіквенсів;

$1 - (1-p_{iL})^n$ – імовірність хоч раз ідентифікувати конкретну лінію L у n сіквенсах;

$p_{iL} = k_{iL}/N_i$ – вага лінії у потоці зразків, що представляє собою частку лінії в реально спостереженій сумі країн.

За таким розрахунком сума за всіма лініями, що мають відмінну від нуля частоту, дає очікуване покриття набору різноманіття. Для аналізу мінімальної кількості секвенувань, яке досягає частки від спостереженого різноманіття, використовували наступний підхід:

$$n_q = min\{n \in N: E[S_i(n)] \geq qS_i\}, \qquad q \in \{0{,}90; 0{,}95; 0{,}99\}$$

Запропонована модель не вимагає однакового обсягу вибірки ліній між різними країнами, оскільки здійснює індивідуальне оцінювання порогу ідентифікації відносно конкретного профілю і часток ліній у потоці. Окрім цього, модель спирається виключно на фактичні спостереження ідентифікації ліній, які ілюструють динаміку змін профілю ліній *B. pandemicum*, а отже умовно не залежать від конкретного показника швидкості еволюційного процесу.

Оцінка маржинальної віддачі секвенування.

Маржинальну віддачу визначали як:

$$MR = \frac{dS}{dn},$$

де (S) – очікувана кількість унікальних ліній.

Оцінку проводили чисельно для кожної країни, нормуючи результат на 1000 додаткових сіквенсів.

Дизайн-ефект та сценарний аналіз. Оскільки реальні національні програми секвенування часто містять кластеризований відбір (спалахи, лікарні, таргетовані групи), у моделювання включено сценарний аналіз впливу дизайн-ефекту.

Використовували мультиплікативні коефіцієнти: 1,2, 1,5, 2,0, 3,0, які моделюють збільшення дисперсії та зменшення ефективного розміру вибірки.

Програмне забезпечення та статистичні методи. Обробку даних здійснено з використанням Python 3.11, бібліотек pandas, NumPy, openpyxl.

Для оцінки кореляцій застосовували ранговий коефіцієнт Спірмена, оскільки розподіли були нелінійними та неоднорідними.

Результати досліджень та їх обговорення

Описові статистики профілів країн. За результатами аналізу даних 176 країн сформовано глобальний медіанний профіль ліній *B. pandemicum*. Описові статистики параметрів різноманіття наведено у таблиці 1.

Медіанні значення характеристик профілю вказували на помірне різноманіття ліній (*HMM* ≈ 1,27), помірну концентрованість (*HHI* ≈ 0,38), нерівномірний розподіл часток (*J* ≈ 0,61).

Розподіли *HHI*, *HMM* та J характеризувалися значним розкидом і асиметрією, що відображало різну глибину генетичного моніторингу між країнами. Згідно з отриманими результатами, у кластері країн із мінімальними обсягами секвенувань спостерігався високий *HHI* та низькі значення *HMM*, що свідчило про штучну монопереважальність, зумовлену малим розміром вибірки.

Правостороннє зміщення *HHI* демонструє вплив джерел з низькою кількістю секвенувань, які формували профілі з вузьким переліком іденти-

Таблиця 1

Описові статистики параметрів модельного світового профіля ліній *Betacoronavirus pandemicum* згідно з базою даних Pango

Table 1

Descriptive statistics of the parameters of the model global lineage profile of *Betacoronavirus pandemicum* according to the Pango database

Параметр	Індекс Герфіндаля-Гіршмана, *HHI*	Ентропія Шеннона, *H*	*H* Міллера–Медоу, *HMM*	Нормалізація Пілоу, *J*
Середнє	0,4538	1,4128	1,4262	0,5930
Медіана	0,3844	1,2597	1,2702	0,6116
Мінімум	0,0327	0,0000	0,0000	0,0679
Максимум	1,0000	4,3359	4,3361	0,9977
Розкид	0,9673	4,3359	4,3361	0,9298
Стандартне відхилення	0,3058	1,0632	1,0599	0,1887

фікованих ліній та зміщували глобальну оцінку. Кластер країн з найбільшим значенням *HHI* значною мірою перекривався з найнижчими значенням *HMM*, що відображало прогнозовану сильну обернену асоціацію концентрованості та різноманіття. Діапазон *HMM* охопив спектр від майже монопереважальних до високо деконцентрованих профілей, що характеризувалося зміною ефективної кількості ліній від ≈ 3,6 за медіанним до ≈76 за максимальним значенням. Найвищі значення даного показника формувалися переважно за рахунок масштабних та тривалих програм секвенування, здійснених окремими країнами, обсяг яких охоплював від мільйонів до сотень тисяч наданих сіквенсів. Кластер країн з високим *HMM* був очікувано асоційованим з низьким рівнем *HHI* ≈ 0,03–0,09 та помірним значенням *J* ≈ *0,57–0,79* з великою кількістю ідентифікованих ліній. Встановлене лівостороннє зміщення *J* також формувалося за рахунок нерівномірних профілей з малим обсягом вибірки. Максимальне значення даного показника спостерігалося у країнах з низьким обсягом вибірки та небагатьма лініями. Аналіз полярної вибірки продемонстрував конфігурації профілів з переважанням однієї або кількох ліній з охопленням ~ 500 сіквенсів на країну та виявленим спектром у 2–4 лінії. Комбінація високого рівня *HHI* з низьким *HMM* та *J* зазвичай відповідали монопереважальному профілю, що не завжди було зумовлено критично малим обсягом секвенувань. Практично це свідчило, що високі значення *HMM* на тлі низького *HHI* є надійним маркером системного розгорнутого геномного нагляду. Натомість високе значення *J*, особливо з обмеженим обсягом вибірки, відображає неповноту охоплення, а не реальну рівномірність. Такими чином, всі три показники доцільно використовувати у комбінації для коректної оцінки різноманіття.

Регресійні моделі залежності якості профілю від валової кількості секвенувань.

Для здійснення аналізу теоретичного впливу обсягу секвенувань на якість формування профілів країн здійснено регресійну апроксимацію, що

було спрямовано на пошук оптимальної моделі взаємозв'язку даного показника з характеристиками профілів з використанням 12 моделей.

Згідно з отриманими результатами, залежність між валовим обсягом секвенувань та числом ідентифікованих ліній описується кореневою моделлю з високою пояснювальною здатністю ($R^2 \approx 0{,}96$) (табл. 2). У цій зв'язці спостерігалася сублінійна крива накопичення, де приріст нових ліній був різким на малих обсягах секвенувань, коли перші 50–200 сіквенсів дають різке зростання виявлених ліній, і поступово сповільнювався через перехід до ідентифікації більш рідкісних варіантів *B. pandemicum*.

Таблиця 2

Оптимальні моделі зв'язку валового обсягу секвенувань з параметрами профілів 176 країн за моніторингом динаміки ліній *Betacoronavirus pandemicum*

Table 2

Optimal models linking total sequencing volume with profile parameters across 176 countries in the monitoring of *Betacoronavirus pandemicum* lineage dynamics

Параметр	Модель	R^2	Помилка	Рівняння
Кількість секвенувань ~ *кількість ідентифікованих ліній*	Коренева	0,9639	33,807205	$k = -7{,}49 + 0{,}74 \times \sqrt{n}$
Кількість секвенувань ~ *HMM*	Квадратичний лог-поліном	0,71013	0,567696	$H_{MM} = -0{,}01 + 0{,}15\, ln(n) + 0{,}01(ln(n))^2$
Кількість секвенувань ~ *HHI*	Квадратичний лог-поліном	0,518354	0,210271	$HHI = 0{,}99 - 0{,}10\, ln(n) + 0{,}0025(ln(n))^2$
Кількість секвенувань ~ *J*	Обернена	0,094319	0,179037	$J = 0{,}58 + \frac{1{,}38}{n}$

Зв'язок із різноманіттям ~ *HMM* найкраще описувався квадратичною лог-поліноміальною моделлю, що демонструвало зростання різноманіття зі збільшенням масштабу спостережень відповідно до *ln* кількості секвенувань ($R^2 \approx 0{,}71$). У такому сценарії нарощування валової кількості секвенувань подовжує хвіст низькочастотних варіантів, розширюючи спектр ідентифіковних ліній та частково, сублінійно відносно кількості сіквенсів, вирівнюючи рівномірність.

Зв'язок із концентрованістю (*HHI*) також найкраще описувався квадратичною лог-поліноміальною моделлю ($R^2 \approx 0{,}52$), однак з меншою пояснювальною силою. Взаємозв'язок вказує, що збільшення валового обсягу секвенувань розширює «довгий хвіст» рідкісних варіантів, ріст різноманіття є нелінійним і сповільнюється на великих вибірках.

Зі збільшенням кількості секвенувань переважання окремих ліній слабшає, а розподіл часток стає менш нерівномірним.

Жодна з моделей не продемонструвала прийнятної пояснювальної сили для рівномірності Пілоу (J) ($R^2 < 0,10$), що свідчить про те, що рівномірність профілю значною мірою залежить від дизайну вибірки, а не від валового обсягу секвенувань.

Отже, незважаючи на істотну роль валової кількості секвенувань у формуванні якісного профілю ліній для переважальності та рівномірності вищу значущість мають інші чинники, що можуть включати в себе географічні та часові неоднорідності епідемічного процесу, селективні або логістичні характеристики зразків. У інших випадках наявність сублінійної динаміки описує розширення спектру нових ліній, що виявляються, на малих обсягах секвенувать із подальшим сповільненням, оскільки рідкісні варіанти трапляються дедалі рідше.

Враховуючи багатофакторність впливів, що зумовлюють якість профілів, до моделі включено часову компоненту. Відповідно до специфіки даних, що фіксує база PangoLin, зокрема час отримання першого та останнього зразку в межах певної країни та певної лінії, в процесі моделювання враховували відсутність прямого зв'язку між цими даними та фактичними часовими межами циркуляції збудника і необхідність досягнення порогу аналітичної чутливості стратегії генетичного моніторингу. Згідно з цим, часовий параметр задіяно як проксі показник тривалості активного спостереження (*D*) та інтенсивності (*I*), що представляла собою кількість сіквенсів на добу в межах календарного вікна. Для аналізу було здійснено розрахунок рангової кореляції Спірмена з метою виявити монотонну залежність без припущення лінійності та нормального розподілу.

Результати показали узгоджену та статистично переконливу картину кореляції тривалості моніторингового вікна *(D)* з кількістю ідентифікованих ліній ($\rho = 0,885$; $p < 10^{-59}$), високий зв'язок D з різноманіттям ($\rho = 0,824$). Це підтверджує, що довгі періоди моніторингу відображають більш реалістичну еволюцію вірусу і «висвітлюють» рідкісні варіанти (табл. 3).

Таблиця 3

Результати кореляційного аналізу часових метрик моделі з параметрами профілів 176 країн за моніторингом динаміки ліній *Betacoronavirus pandemicum*

Table 3

Results of the correlation analysis between temporal model metrics and profile parameters across 176 countries in the monitoring of *Betacoronavirus pandemicum* lineage dynamics

Характеристики	*D* ~		*I* ~	
	ρ	*p*-значення	*ρ*	*p*-значення
~ обсяг секвенувань	0,885	$< 1,0 \times 10^{-59}$	0,888	$< 1,0 \times 10^{-59}$
~ *HMM*	0,824	$1,7 \times 10^{-44}$	0,768	$2,3 \times 10^{-35}$
~ *HHI*	-0,773	$5,6 \times 10^{-36}$	−0,715	$9,6 \times 10^{-29}$
~ *J*	-0,112	0,16	−0,286	$3,0 \times 10^{-4}$

Для взаємодії *D* з концентрованістю *HHI* виявлено значущий негативний зв'язок (ρ = –0,773), що свідчить про зменшення переважання ліній, пов'язане з розширенням різноманіття та перерозподілом часток кожного окремого варіанту. Рангова кореляція між *D* та *J* не виявила статистично значущого зв'язку, відповідно загальний часовий масштаб сам по собі не є вирішальним чинником формування рівномірності між частками ідентифікованих ліній.

Згідно з цим, довжина вікна спостереження є значним чинником, однак при тлумаченні необхідно здійснювати контроль конфаундерів, які відповідають дійсному контексту, включно з врахуванням стратегії відбору, що може здійснюватися у вигляді рутинного скринінгу або як таргетоване спорадичне розслідування. Ці дві стратегії реалізують свій вплив на часову компоненту моделі за рахунок різного рівня безперервності отримання зразків в межах вікна моніторингу. Саме безперервність може зумовлювати охоплення коротких інтервалів циркуляції, а короткі вікна та фрагментарні періоди скринінгу схильні до демонстрації хибного переважання однієї або групи варіантів, що може сприяти неефективним епідеміологічним та медичним заходам.

За результатами оцінювання наявності монотонного зв'язку між інтенсивністю секвенувань (*I)* та характеристиками профілю виявлено статистично значущу кореляцію між збільшенням *I*, різноманіттям *HMM* (ρ = 0,768) та числом ліній (ρ = 0,888).

Така кореляція демонструє витягування довгого хвоста низькочастотних варіантів *SARS-CoV 2* у відповідності зі збільшенням інтенсивності секвенувань. Асоціація між *I* та концентрованістю *HHI* (ρ = –0,715) свідчить про послаблення переважання окремих ліній на тлі більшого охоплення моніторингу, отже, підвищення інтенсивності зменшує шанси хибної монопереважальності. Значущої кореляції між *I* та *J* виявлено не було.

Результати демонструють, що якість профілю майже однаково зумовлена обома часовими чинниками. Це підкреслює ключову роль обґрунтованого поєднання календарного покриття та щільності випробовувань для отримання репрезентативної картини генетичного різноманіття ліній збудника в межах програми моніторингу. Показник J продемонстрував лише слабкі кореляції з тривалістю спостереження (ρ = –0,286) та інтенсивністю секвенування (ρ = –0,112). Це свідчить про те, що рівномірність розподілу ліній практично не залежить від часових чи кількісних параметрів програми моніторингу та визначається передусім особливостями дизайну відбору зразків.

З огляду на практичну роль типування та моніторингу ліній у стратегії подолання пандемій та її економічну вагу, здійснено аналіз типового кумулятивного діапазону обсягу секвенувань, в межах якого стратегія моніторингу є найбільш ефективною. Охоплення різноманіття моделювали як співвідношення між очікуваною кількістю унікальних ліній у вибірці та імовірністю принаймні однієї ідентифікації заданої лінії з врахуванням її частки у реальному потоці зразків окремої країни. Оцінювали досягнення порогових значень 90, 95 та 99% від спостереженого різноманіття у кожній країні відносно її індивідуального профілю.

Як видно з таблиці 4, досягнення 90% різноманіття потребувало медіанно ~ 239 сіквенсів, тоді як збільшення охоплення лише на наступні 5% (до

95%) вже вимагало додаткових ~ 82 сіквенсів. Подальше розширення до 99% різноманіття потребувало ще ~ 90 сіквенсів, тобто зіставного додаткового обсягу, але давало лише +4% приросту. Таке співвідношення між збільшенням обсягу секвенувань та мінімальним приростом охоплення свідчить, що після ~ 400–500 сіквенсів система входить у зону насичення, де ефективність додаткового секвенування різко знижується.

Таблиця 4

Порогові обсяги секвенувань, необхідні для виявлення 90–99% різноманіття ліній *Betacoronavirus pandemicum*, в межах часового вікна моніторингу

Table 4

Threshold sequencing depths required to detect 90–99% of the diversity of *Betacoronavirus pandemicum* lineages within the monitoring time window

Покриття	Медіана валової кількості секвенувань	Середня валової кількості секвенувань	Медіана сіквенсів на добу	Середнє сіквенсів на добу
90% ліній	239	16161,337	0,275641	9,417386
95% ліній	321	34291,954	0,354254	19,893142
99% ліній	411	64699,840	0,445217	37,348114

Перехід від 95 до 99% охоплення досягався за рахунок інтенсивності секвенування в інтервалі 9,42–37,35 сіквенсів на добу в межах часового вікна моніторингу. Отже, подальше нарощування обсягів для виконання лише мети моніторингу може бути економічно недоцільним.

Це є ключовим результатом створеної моделі.

Окремо оцінювали ефективний обсяг секвенування та інтенсивності для 95% імовірності ідентифікації принаймні однієї рідкісної лінії. Оцінку наведено в таблиці 5. Для ймовірності виявлення 0,95 мінімальні обсяги складають: при частці 1,0% – 299 сіквенсів, при частці 0,1% – 2995 сіквенсів, при частці 0,01% – 29956 сіквенсів.

Згідно з отриманими результатами, вимоги до необхідного обсягу секвенувань зростають обернено пропорційно та кратно до рідкості цільових ліній. Незважаючи на модельний статус, такі результати підкреслюють, що зниження порогу детекції до лінії з малою часткою за підтримкою рівня 95% ймовірності виявлення стає можливим лише за умов суттєвого нарощування пропускної спроможності, подовження вікна моніторингу або суттєвого перегляду дизайну відбору зразків для секвенування.

Оскільки кластеризація у дизайні відбору зразків зменшує ефективний розмір вибірки ідентифікованих ліній за фіксованої кількості сіквенсів і збільшує дисперсію відносно багатономіальної схеми, для адаптації даної моделі є необхідним введення дизайн-ефекту. Однак, оскільки база даних не містила інформації про внутрішні агрегати джерел сіквенсів у країнах-учасниках та часовий розподіл надходження зразків, емпірична оцінка кореляції для кожної країни є неможливою. Для аналізу впливу дизайн-ефекту задіяно сценарний аналіз чутливості з фіксованими множинниками у 1,2, 1,5, 2,0 та 3,0. За

Таблиця 5

Порогова інтенсивність секвенувань, що є необхідною для ідентифікації рідкісних ліній *Betacoronavirus pandemicum* протягом часового вікна моніторингу

Table 5

Threshold sequencing intensity required for the detection of rare *Betacoronavirus pandemicum* lineages within the monitoring time window

Частка лінії в потоці секвенування	Мінімальна валова кількість сіквенсів за інтервал спостережень	Мінімальна середня добова інтенсивність сіквенсів для виявлення за 7 діб	Мінімальна середня добова інтенсивність сіквенсів для виявлення за 14 діб	Мінімальна середня добова інтенсивність сіквенсів для виявлення за 30 діб
1,0% (0,0100)	299	42,714	21,357	9,967
0,5% (0,0050)	598	85,429	42,714	19,933
0,1% (0,0010)	2995	427,857	213,929	99,833
0,05% (0,0005)	5990	855,714	427,857	199,667
0,01% (0,0001)	29956	4279,429	2139,714	998,533

умов подібності зразків у межах кластерів зростає міжкластерна кореляція, і додаткові сіквенси, що походять с того самого кластера, додають меншу кількість нових ліній, а отже цільовий загальний обсяг та добова інтенсивність у такому сценарії масштабуються пропорційно внеску дизайн-ефекту. Найбільш відчутний вплив цієї поправки відчувається на рівні 95% ймовірності хоча б одноразової ідентифікації лінії з часткою 0,1%, що в межах 14-денного вікна зростає з 428 сіквенсів на добу до рівня 500–800 сіквенсів на добу. Це вчергове підкреслює гостру необхідність гармонізації дизайну відбору зразків в межах програми моніторингу.

Для встановлення верхньої раціональної межі секвенувань здійснено аналіз маржинальної віддачі, яка формує очікуваний приріст кількості унікальних ліній при збільшенні обсягу секвенувань на одиницю. Аналіз здійснювали за допомогою оцінювання миттєвої маржинальної віддачі на базі формування похідної з результатів оцінювання загальних порогових рівнів секвенування та перераховували у вигляді віддачі як кількості нових ліній, що вдається ідентифікувати на кожні 1000 додаткових секвенувань.

Агреговано для 176 країн (див. Додаток) отримано медіанне значення близько 0,64 нових ліній на кожні 1000 секвенувань. Таким чином, для типової країни додаткові тисяча сіквенсів на рівні ідентифікації, що наближалася до 09.10.2024, приносили менш ніж одну нову унікальну лінію. Водночас, деякі країни що не входять до 95 перцентилю, все ще здатні ідентифікувати близько 23,55 нових ліній при збільшенні валової кількості секвенувань на 1000.

Враховуючи результати, світову модель молекулярно-генетичного моніторингу слід коригувати в режимі реального часу з встановленням динамічного критерію ефективності, який має враховувати конкретну інтенсивність

детекції унікальних ліній для конкретної країни, та її вклад у загальну картину змін генетичної структури збудника. Такий підхід дозволить оптимізувати витрати з встановленням пріоритетів інвестування.

За результатами ретроспективного моделювання світового моніторингу динаміки ліній *B. pandemicum* встановлено, що для ефективної протидії ймовірним наступним пандемічним інцидентам необхідною є оптимізація глобальної стратегії. У першу чергу слід врахувати необхідність встановлення обґрунтованих метрик контролю якості програми. Узгоджений потоковий аналіз показників концентрованості, різноманітності та рівномірності профілів з обов'язковими поправками на дизайн-ефект і контекст відбору має високий потенціал підвищення ефективності за допомогою впровадження відповідних коригувань як у програму моніторингу, так і у її матеріальне забезпечення. Моніторинг таких метрик зумовлює можливість введення тригерів реагування, наприклад, раптове зростання переважання, стале падіння різноманітності за незмінної інтенсивності секвенувань, поява і короткочасна персистентність низькочастотних кластерів у різних регіонах тощо. Такі тригери мають істотне значення у питанні зміни пріоритезації ресурсів.

Необхідним є не лише безперервне календарне покриття, а й його поєднання з достатньою інтенсивністю секвенувань, яке має спиратися на аналіз індивідуального профілю. Такий підхід дозволить ідентифікувати рідкісні варіанти на ранніх етапах та знизити хибне переважання окремих ліній, що має суттєвий вплив на вакцинальну кампанію. Окрім цього, стратегія моніторингу повинна враховувати формування кривої накопичення, яке на ранніх стадіях моніторингу не дозволяє ефективно ідентифікувати як спектр ліній так і їх переважальність, що формує необхідність максимально швидкого виходу на оптимальні рівні валової кількості сіквенсів, оскільки основну прогностичну цінність мають не миттєві показники, а динаміка метрик, що має здатність відрізняти еволюційну новизну від шумових ефектів.

Додаток до статті можна переглянути в електронній версії журналу за посиланням: http://mbt.onu.edu.ua/article/view/345238

K. S. Onyshchenko[1], O. Yu. Zinchenko[2], B. S. Zhukov[3]
[1]National Pirogov Memorial Medical University,
56 Pyrohova St, Vinnytsia 21018, Ukraine,
e-mail: cat.medichka@gmail.com
[2]Odesa I. I. Mechnikov National University,
2 Zmiienka Vsevoloda St, Odesa, 65082, Ukraine,
e-mail: ozinchenko@onu.edu.ua
[3] Odesa National University of Technology,
112 Kanatna St, Odesa 65039, Ukraine,
e-mail: brs.alchemist@gmail.com

RETROSPECTIVE MATHEMATICAL MODELING OF GLOBAL *BETACORONAVIRUS PANDEMICUM* LINEAGES DYNAMIC MONITORING AS AN INSTRUMENT FOR THE SURVEILLANCE OF EMERGING MEDICAL THREATS

Summary

Aim. *To assess the parameters of SARS-CoV-2 (Betacoronavirus pandemicum) lineage profiles and to determine the relationships between sequencing depth, temporal monitoring characteristics, and lineage diversity.* ***Methods.*** *A dataset 11.33 million sequences and 2,703 B. pandemicum lineages from 176 countries were analyzed using the Pango nomenclature. Concentration (HHI), diversity (H, HMM), evenness (J), observation duration (D), and sequencing intensity (I) indices were calculated. Correlation and regression analyses were performed using appropriate Python libraries.* ***Results.*** *Strong associations were identified between monitoring duration, sequencing intensity, and the number of detected lineages. Threshold sequencing depths required to reach 90%, 95%, and 99% lineage diversity coverage were ~239, 321, and 411 sequences, respectively. After ~400–500 sequences, the efficiency of sequencing decreased, reflected by a minimal increase in newly detected lineages. The marginal yield fell below one new lineage per 1,000 additional sequences.* ***Conclusions.*** *Temporal parameters and sequencing intensity determine the effectiveness of genomic surveillance, whereas profile evenness depends mainly on sampling design. The proposed model enables determination of sequencing thresholds and optimization of monitoring strategies.*

Key words: COVID-19, Betacoronavirus pandemicum, design-model, sequencing, diversity, surveillance.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Brainard J., Jones N., Harrison F., et al.* Super-spreaders of novel coronaviruses that cause SARS, MERS and COVID-19: a systematic review // Annals of Epidemiology. – 2023. – V. 82. – P. 66–76. – https://doi.org/10.1016/j.annepidem.2023.03.009
2. *Colquhoun R., Jackson B., O'Toole A., et al.* SCORPIO: a utility for defining and classifying mutation constellations of virus genomes // Bioinformatics. – 2023. – V. 39(10). – 4 p. – https://doi.org/10.1093/bioinformatics/btad575
3. *Du S., Tong X., Lai A., et al.* Highly host-linked viromes in the built environment possess habitat-dependent diversity and functions for potential

virus-host coevolution // Nature Communications. – 2023. – V. 14(2676). – 16 p. – https://doi.org/10.1038/s41467-023-38400-0
4. *Herrera A., Riera R., Rodríguez R.* Alpha species diversity measured by Shannon's H-index: some misunderstandings and underexplored traits, and its key role in exploring the trophodynamic stability of dynamic multiscapes // Ecological Indicators. – 2023. – V. 156. – 10 p. – https://doi.org/10.1016/j.ecolind.2023.111118
5. *Kvalseth T.* Measurement of market (industry) concentration based on value validity // PLoS ONE. – 2022. – V. 17(7). – 24 p. – https://doi.org/10.1371/journal.pone.0264613
6. *Li Q., Shah T., Wang B., et al.* Cross-species transmission, evolution and zoonotic potential of coronaviruses // Frontiers in Cellular and Infection Microbiology. – 2023. – 13 p. – https://doi.org/10.3389/fcimb.2022.1081370
7. *Liu D., Chen C., Chen D., et al.* Mouse models susceptible to HCoV-229E and HCoV-NL63 and cross protection from challenge with SARS-CoV-2 // Proceedings of the National Academy of Sciences of the USA. – 2023. – V. 120(4). – 12 p. – https://doi.org/10.1073/pnas.2202820120
8. *Lopez-Cortes G., Palacios-Perez M., Hernandez-Aguilar M.* Human Coronavirus Cell Receptors Provide Challenging Therapeutic Targets // Vaccines. – 2023. – V. 11(1). – 23 p. – https://doi.org/10.3390/vaccines11010174
9. *O'Toole A., Pybus O., Abram M., et al.* Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences // BMC Genomics. – 2022. – V. 23. – 121 p. – https://doi.org/10.1186/s12864-022-08358-2
10. *O'Toole A., Scher E., Underwood A., et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool // Virus Evolution. – 2021. – V. 7(2). – P. 1–9. – https://doi.org/10.1093/ve/veab064
11. *Rambaut A., Holmes E., O'Toole A., et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology // Nature Microbiology. – 2020. – V. 5. – P. 1403–1407. – https://doi.org/10.1038/s41564-020-0770-5
12. *Roswell M., Dushoff J., Winfree R.* A conceptual guide to measuring species diversity // Oikos. – 2021. – V. 130(3). – P. 321–487. – https://doi.org/10.1111/oik.07202
13. *Spiegel Y.* The Herfindahl-Hirschman Index and the Distribution of Social Surplus // The Journal of Industrial Economics. – 2021. – V. 69(3). – P. 561–594. – https://doi.org/10.1111/joie.12253
14. *Zumla A., Peiris M., Memish Z., et al.* Anticipating a MERS-like coronavirus as a potential pandemic threat // Lancet. – 2024. – V. 4(403). – P. 1729–1731. – https://doi.org/10.1016/S0140-6736(24)00641-X

REFERENCES

1. Brainard J, Jones N, Harrison F, et al. Super-spreaders of novel coronaviruses that cause SARS, MERS and COVID-19: a systematic review.*Ann Epidemiol*. 2023;82:66–76. https://doi.org/10.1016/j.annepidem.2023.03.009

2. Colquhoun R, Jackson B, O'Toole A, et al. SCORPIO: a utility for defining and classifying mutation constellations of virus genomes. *Bioinformatics*. 2023;39(10):1–4. https://doi.org/10.1093/bioinformatics/btad575
3. Du S, Tong X, Lai A, et al. Highly host-linked viromes in the built environment possess habitat-dependent diversity and functions for potential virus-host coevolution. *Nat Commun*. 2023;14(2676):1–16. https://doi.org/10.1038/s41467-023-38400-0
4. Herrera A, Riera R, Rodríguez R. Alpha species diversity measured by Shannon's H-index: some misunderstandings and underexplored traits, and its key role in exploring the trophodynamic stability of dynamic multiscapes. *Ecol Indic*. 2023;156:111118. https://doi.org/10.1016/j.ecolind.2023.111118
5. Kvalseth T. Measurement of market (industry) concentration based on value validity. *PLoS One*. 2022;17(7):e0264613. https://doi.org/10.1371/journal. pone.0264613
6. Li Q, Shah T, Wang B, et al. Cross-species transmission, evolution and zoonotic potential of coronaviruses. *Front Cell Infect Microbiol*. 2023;13:1081370. https://doi.org/10.3389/fcimb.2022.1081370
7. Liu D, Chen C, Chen D, et al. Mouse models susceptible to HCoV-229E and HCoV-NL63 and cross protection from challenge with SARS-CoV-2. *Proc Natl Acad Sci USA*. 2023;120(4):e2202820120. https://doi.org/10.1073/ pnas.2202820120
8. Lopez-Cortes G, Palacios-Perez M, Hernandez-Aguilar M. Human coronavirus cell receptors provide challenging therapeutic targets. *Vaccines*. 2023;11(1):1–23. https://doi.org/10.3390/vaccines11010174
9. O'Toole A, Pybus O, Abram M, et al. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. BMC Genomics. 2022;23(1):121. https://doi.org/10.1186/s12864-022-08358-2
10. O'Toole A, Scher E, Underwood A, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol*. 2021;7(2):veab064. https://doi.org/10.1093/ve/veab064
11. Rambaut A, Holmes E, O'Toole A, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. 2020;5:1403–1407. https://doi.org/10.1038/s41564-020-0770-5
12. Roswell M, Dushoff J, Winfree R. A conceptual guide to measuring species diversity. *Oikos*. 2021;130(3):321–487. https://doi.org/10.1111/oik.07202
13. Spiegel Y. The Herfindahl-Hirschman Index and the distribution of social surplus. *J Ind Econ*. 2021;69(3):561–594. https://doi.org/10.1111/joie.12253
14. Zumla A, Peiris M, Memish Z, et al. Anticipating a MERS-like coronavirus as a potential pandemic threat. *Lancet*. 2024;403:1729–1731. https://doi.org/10.1016/ S0140-6736(24)00641-X

Стаття надійшла до редакції 01.12.2025 р.