КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ВИСОКИХ ТЕХНОЛОГІЙ

Завідувач кафедри молекулярної біотехнології та біоінформатики
к. б. н., доцент Нипорко О. Ю.
Протокол №____ засідання кафедри
від “____” _____2023 р.

# ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ВИЯВЛЕННЯ ТА КІЛЬКІСТНОЇ ОЦІНКИ ЛІНІЙ SARS-COV-2 У ЗРАЗКАХ СТІЧНИХ ВОД

Випускна кваліфікаційна робота магістра
студента спеціальності 091
Біологія
ОП «Біоінфоматика та структурна біологія»
Тищенко Богдана Юрійовича

Науковий керівник від кафедри
професор кафедри молекулярної біотехнології
та біоінформатики
д. б. н. Солдаткін О.П.

Робота виконана у Лабораторії Мангула
при Університеті Південної Каліфорнії
під керівництвом Мангул С.

Оцінка захисту роботи

_____

Київ – 2023 р.

# АНОТАЦІЯ

Тищенко Б.Ю. Порівняльний аналіз методів виявлення та кількісної оцінки ліній SARS-CoV-2 у зразках стічних вод. - Випускна кваліфікаційна робота магістра за спеціальністю 091 Біологія ОП «Біоінформатика та структурна біологія».

У роботі проведено оцінку і порівняння ефективності методів детекції ліній на даних SARS-COV-2, включаючи CliqueSNV, PredictHaplo, aBayesQR та gromstole. Згенеровано масивні штучні набори даних рідів зі зразками 11ти варіантів SARS-COV-2, які можуть слугувати основою для оцінки якості та надійності алгоритмів виявлення варіантів і визначення їх концентрацій. Отримані результати вказують на проблеми в роботі деяких методів з наборами даних повного геному, а також на можливості покращення результатів за допомогою налаштування параметрів алгоритмів. Отримані результати можуть бути використані для вдосконалення існуючих та розробки нових методів детекції варіантів SARS-COV-2, а також для визначення найбільш ефективних підходів для аналізу даних про варіанти цього вірусу.

Ключові слова: порівняльний аналіз, детекція ліній, SARS-COV-2, стічні води, CliqueSNV, PredictHaplo, aBayesQR, gromstole.

## ABSTRACT

Tyshchenko B.Yu. Benchmarking of methods for detecting and quantitatively assessing SARS-CoV-2 strains in wastewater samples. - Master's thesis for the specialty 091 Biology OP "Bioinformatics and Structural Biology".

This study evaluates and compares the effectiveness of strain detection methods on SARS-COV-2 data, including CliqueSNV, PredictHaplo, aBayesQR, and gromstole. Large synthetic data sets of strains with samples of 11 variants of SARS-CoV-2 have been generated, which can serve as a basis for evaluating the quality and reliability of strain detection algorithms and determining their concentrations. The results obtained indicate problems in the work of some methods with full genome data sets, as well as the possibility of improving results by adjusting algorithm parameters. The results can be used to improve existing and develop new methods for detecting SARS-CoV-2 strains, as well as to identify the most effective approaches for analyzing data on variants of this virus.

Keywords: benchmarking, strain detection, SARS-CoV-2, wastewater, CliqueSNV, PredictHaplo, aBayesQR, gromstole.

# ЗМІСТ

# ВСТУП

Науково-дослідницька робота проходила віддалено, на базі Лабораторії Мангула при Університеті Південної Каліфорнії. Об’єктом дослідження були методи детекції варіантів вірусів у метагеномних даних.

## 1.1 Мотивація

Вірус SARS-CoV-2 нещодавно став прикладом того, як небезпечні віруси можуть протягом довгого часу поширюватися, розвиватися та мутувати, маючи потенціал до породження більш небезпечних варіантів. Через це, потреба в економічно вигідному та дієвому способі виявлення присутності різних ліній та їх кількості в популяції вийшла за рамки термінової. Було доведено, що геномний нагляд за стічними водами виявляє патогени в каналізаційних сховищах, які є хорошим відображенням колективних відходів, таких як фекалії та сеча. Геномний нагляд за стічними водами має численні привабливі переваги перед клінічним моніторингом, наприклад, моніторинг стічних вод враховує випадки від недостатньо охоплених та вразливих груп населення, тоді як поточний клінічний моніторинг цього не робить. Крім того, подібні методи можна застосовувати для моніторингу інших вірусних захворювань, таких як грип, тощо, без значного збільшення вартості моніторингу.[1]

Результати, представлені в цій роботі, допоможуть дослідникам зробити обґрунтований вибір найбільш прийнятного та точного методу виявлення SARS-CoV-2 у стічних водах, який може задовольнити потреби

конкретного проекту. Коли дослідник обирає детектор вірусу на підставі таких суворих доказів, виявлення генеалогії SARS-CoV-2, для виявлення мутацій і гаплотипів SARS-CoV-2, у стічних водах можна справді масштабувати і підійти до цього економічно ефективно. Окрім цього, ці методи можна застосовувати для подальшого передбачення нових відхилень до того, як вони стануть поширеними в загальній популяції - це висновок зроблений у статті[1].

## 1.2 Труднощі моніторингу стічних вод і шлях до їх вирішення

Геномний нагляд за стічними водами створює технічні проблеми, включаючи погане охоплення/якість секвенування через низькі концентрації та низьку якість РНК та наявність інгібіторів ПЛР, які можуть заважати підготовці бібліотеки для складних зразків навколишнього середовища. Ці проблеми можуть призвести до неповного покриття геному та різної глибини секвенування. Крім того, поточні інструменти для класифікації ліній SARS-CoV-2 були в основному розроблені для клінічних зразків, що містять один домінантний варіант, і не можуть оцінити відносну кількість кількох ліній у зразках із сумішшю вірусів, таких як зразки стічних вод.

Незважаючи на ці проблеми, нещодавно, була проведена робота з використанням різних інструментів біоінформатики, щоб запропонувати можливі способи моніторингу варіантів SARS-CoV-2 і прогнозувати поширеність варіантів за допомогою обчислювальних підходів. Однак ці підходи створюють нові обмеження: 1. Усі підходи зіткнулися з труднощами визначення ліній із низькою концентрацією, особливо при використанні даних з одного місця збору стічних вод. 2. Рідкісні лінії спостережень не

були виявлені в клінічних зразках, що означає, що рідкісні лінії можна пропустити. 3. Відмінності в протоколах секвенування призводять до різних результатів. Хоча для отримання достатньої інформації часто необхідне повногеномне, а не спайкове, секвенування білка, воно дає неоднозначні результати порівняно з спайковим секвенуванням білка. Крім того, глибина охоплення генома SARS-CoV-2, секвенованого зі зразків стічних вод, як правило, нерівномірна. Загалом прогнозування SARS-CoV-2 за допомогою інструментів біоінформатики є складним через відсутність вказівок, отриманих на основі систематичного порівняння всіх нещодавно розроблених методів прогнозування SARS-CoV-2 у стічних водах.

Щоб вирішити ці ключові проблеми, які наразі обмежують дослідження стічних вод SARS-CoV-2, ми плануємо провести масштабний порівняльний аналіз інструментів біоінформатики з використанням різних джерел даних. Зокрема, для створення ми завантажили 5 різних вірусних штамів із GISAID (альфа, бета, дельта, гамма та омікрон). Ми створимо in-silico суміш різної однакової кількості штамів і згенеруємо in-silico Illumina, PacBio та Nanopore зчитування, що дозволить нам порівняти інструменти щодо здатності визначати штами у зразку за допомогою SimSeq. Загалом 26 методів будуть перевірені, щоб дати еталон для моніторингу на основі геному стічних вод.

Щоб вирішити ці ключові проблеми, які наразі обмежують дослідження стічних вод SARS-CoV-2, ми плануємо провести масштабний порівняльний аналіз інструментів біоінформатики з використанням різних джерел даних. Зокрема, для створення ми завантажили 5 різних вірусних штамів із GISAID (альфа, бета, дельта, гамма та омікрон). Загалом 5 методів будуть перевірені, щоб дати еталон для моніторингу на основі геному стічних вод.

# 2 Розділ 1. Огляд

## 2.1 Симуляція рідів секвенування

Симуляція зчитувань необхідна для оцінки надійності інструментів NGS, оскільки вона дозволяє генерувати контрольовані дані, які тісно імітують реальні дані секвенування. У контексті виявлення варіантів SARS-CoV-2 у відходах, ці симуляції можуть допомогти у прогнозуванні та оцінці продуктивності алгоритмів виявлення варіантів за різних умов.

### 2.1.1 SimSeq

SimSeq працює, імітуючи процес секвенування у контрольованому обчислювальному середовищі. Процес починається з визначеного користувачем референтного геному або послідовності, які можуть варіюватися від геному цілого організму до одного певного гену, або конкретного варіанту.

SimSeq використовує так звані профілі помилок для симуляції мутацій та помилок, що з’являються в процесі секвенування. Ці помилки включають точкові мутації, вставки та видалення. Їх додавання SimSeq зазвичай обумовлюється параметрами, визначеними користувачем, або емпіричними профілями помилок, створених на базі існуючих наборів даних.

Інструмент також дозволяє вказати покриття та глибину зчитування - кількість разів, коли кожна база у референтній послідовності секвенується, та скільки зчитувань покриває певну основу відповідно. Ці фактори критично важливі для точного виявлення варіантів: вище покриття та глибина зчитування зазвичай призводять до більш точного виявлення.

Використання SimSeq вимагає надання конкретних вхідних параметрів, таких як референтна послідовність, модель помилок та бажане покриття та глибина зчитування. Референтну послідовність можна вибрати з урахуванням досліджуваного об'єкта - для нашого дослідження послідовностями різних варіантів SARS-CoV-2 служать як референт.

#### **2.1.1.1** Переваги та обмеження

Однією з ключових переваг SimSeq є його можливості налаштування. Користувачі мають значний контроль над параметрами симуляції, що дозволяє їм тісно імітувати реальні ситуації секвенування. Інструмент особливо цінний при випробуванні нових інструментів або алгоритмів, а

також при перевірці того, як добре існуючі інструменти працюють за різних умов.

Однак у SimSeq також є обмеження. Хоча модель помилок намагається відтворити реальні помилки секвенування, вона може не відображати усі складності та нюанси цих помилок. Крім того, вона не враховує всі можливі біологічні та технічні фактори, які можуть впливати на реальне секвенування, такі як наявність забруднювачів або упередженості, що вводяться під час підготовки бібліотеки.[2]

### 2.1.2 SWAMPy

Загальний процес роботи SWAMPy можна побачити на Рисунку 2.1. Чотири базові етапи нашого програмного пайплайну детально описані нижче:

1. Створення початкової популяції ампліконів.

   Програмне забезпечення передбачає, що користувач надає набір геномів SARS-CoV-2, і обирає набір праймерів з підтримуваних програмою наборів праймерів. За замовчуванням вибрано набір праймерів ARTIC версії 1. На основі цього вибору амплікони вилучаються з кожного геному. Спочатку використовується Bowtie 2 для вирівнювання праймерів (прямий та зворотний комплементарний) з кожним вірусним геномом, щоб визначити позиції зв'язування праймерів на початкових геномах. Потім геноми розрізаються для зв'язування праймерів, щоб отримати окремі амплікони кожного геному, включаючи послідовності праймерів. Під час кроку вирівнювання деякі праймери можуть не добре вирівнятися з вірусним геномом, і в таких випадках відповідний амплікон не виробляється. Це суворе покарання за невідповідність.

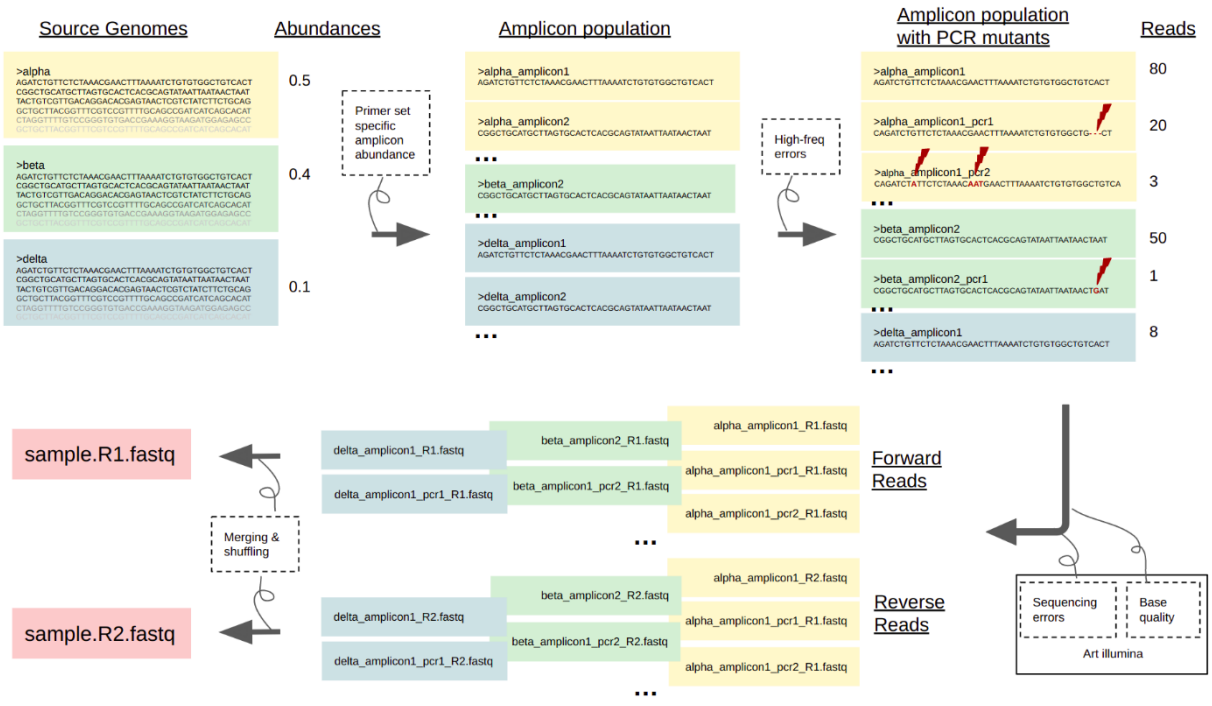



*Рисунок 2.1* [3]

2. Моделювання фрагментів ДНК (копій) для кожного амплікону.

Для моделювання копій на амплікон та геном ми пропонуємо два варіанти комбінованої моделі: поліноміальної та моделі Дерихле.

Користувач повинен вказати три параметри: загальне цільове число рідів, вектор довжин геномів і параметр Дерихле. Вибір цього параметра приблизно відповідає оцінці якості зразка: вище значення відповідає високоякісним зразкам і нижче значення відповідає низькоякісним зразкам.

SWAMPy надає емпірично отриману оцінку пропорцій ампліконів, індексовану по ампліконам.

У "Моделі 1" пропорції ампліконів вважаються однаковими для всіх геномів.

"Модель 2" передбачає різні пропорції ампліконів для різних геномів.

Одним з нюансів у цьому процесі є те, що загальна кількість рідів може бути менше потрібної, якщо деякі амплікони втрачаються через мутації на місцях прив'язки праймерів.

3. Симуляція помилок високої частоти шляхом мутацій ампліконів у популяції.

1) Вибірка високочастотних помилок. Щоб моделювати високочастотні помилки в SWAMPy, спочатку ми створюємо таблицю, яка містить всі високочастотні помилки, що будуть вводитись.

Кількість помилок кожного типу, яка буде вводитись, вибирається з розподілу

$$Poisson(L \times R), \text{де} L = 29903$$

- це довжина референтного геному Wuhan Wuhan-Hu-1, а R - це частота помилок даного типу (вставка, видалення або заміна, кожна або унікальна або повторювана, Рисунок 2.2). Цей розподіл Пуасона наближує розподіл біноміальний, оскільки частоти помилок зазвичай низькі. Частоти помилок можна визначити для кожного з шести типів помилок, зі значеннями за замовчуванням, оціненими на основі реальних експериментів з стічними водами.

Геномна позиція для кожної помилки вибирається випадково без заміни з Wuhan-Hu-1. Для унікальних помилок один із вихідних геномів випадково призначається з вагами вибірки, рівними об'ємам геномів у суміші. Більше того, якщо більше одного амплікону охоплює вже визначену позицію помилки, унікальна помилка призначається лише одному з них. Повторювані помилки призначаються всім вихідним геномам і перекриваючим ампліконам.

2) Застосування симульованих помилок до симульованих ампліконів.

Після того, як ми склали таблицю, що містить всі симульовані помилки, ми обробляємо кожен початковий геном g та кожен амплікон a в популяції ампліконів, яку ми створили раніше. Для кожного a, g:

I. Помилки, які впливають на геном g та амплікон a, вибираються з таблиці помилок.

II. Оскільки симульовані позиції помилок базуються на референтному геномі Wuhan-Hu-1, а варіант амплікону в пробі стічних вод може містити вставки та видалення, послідовності ампліконів вирівнюються до Wuhan-Hu-1 за допомогою Bowtie 2, і визначаються позиції помилок у ампліконі.

III. Для кожної помилки e, визначається кількість прочитань, в яких e присутня, шляхом вибору кількості прочитань n з $Binomial(x_{a,g}, fe)$, де $x_{a,g}$- це загальна кількість прочитань амплікону a для геному g, як описано в пункті 2, а fe - це VAF помилки, як визначено в пункті 3.1.

IV. Для кожної можливої комбінації високочастотних помилок, що впливають на геном g та амплікон a, випадково вибирається кількість прочитань ni, дотримуючись індивідуальних кількостей прочитань помилок. При цьому, не має сенсу симулювати кореляції між помилками на ампліконі, оскільки симуляція спадковості помилок для кожного амплікону є обчислювально дуже важкою, і ми припускаємо, що помилки на ампліконі є незалежними.

V. Нарешті, для кожної комбінації i помилок, які впливають на a і g, створюється нова відповідна модифікована послідовність амплікону.

4. Симуляція рідів секвенування за допомогою ART.

Щоб створити набір симульованих парних прочитань Illumina з кожного амплікону, кожен з яких із заданою кількістю прочитань, використовується

програма ART: режим парних ампліконів ART, а також прапори noALN і maskN. Ці налаштування створюють парні прочитання по 150 п.о., і коректно переписують будь-які символи "N", що з'являються в ампліконі. Використовується набір стандартних коефіцієнтів помилок і профілів якості, налаштовані для секвенатора Illumina MiSeq V3, хоча пакет ART має налаштування для інших платформ і довжин прочитань. [4]

Нарешті, ми використовуємо користувацький скрипт на основі універсальних утиліт bash для з'єднання всіх файлів прочитань FASTQ і перетасування їх порядку, щоб уникнути потенційних упереджень.[3]

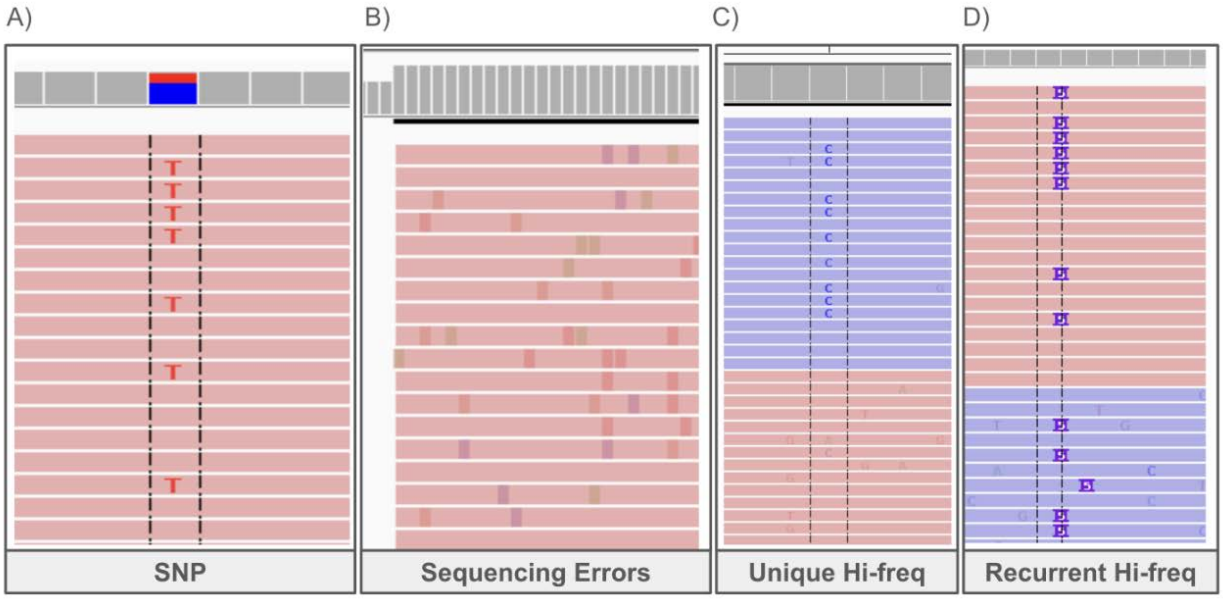



*Рисунок 2.2* Зображення IGV (22) прочитань, симульованих за допомогою SWAMPy. Червоні та сині смуги - це відповідно прямі та зворотні прочитання в режимі "додаткові вирівнювальні зв'язки" IGV. A) Реальний SNP між різними варіантами SARS-CoV-2. B) Помилки секвенування, додані ART. Це зображення з кінця прочитання, де щільність помилок секвенування вища. C) Унікальна помилка високої частоти. Вона з'являється тільки в одному напрямку прочитання і, незважаючи на те, що цей приклад був обраний у регіоні перекриття ампліконів, лише один з ампліконів містить помилку. D) Повторювана помилка високої частоти, яка з'являється в обох напрямках прочитання, і в обох ампліконах, що покривають обраний регіон.[3] [4]

## 2.2 Методи геномної реконструкції

### 2.2.1 CliqueSNV

Гаплотипи - це комбінації алелей (варіації гену), які знаходяться на одній хромосомі і зазвичай успадковуються разом. Реконструкція гаплотипів, також відома як фазування гаплотипів, - це критичний процес в геноміці. Процес включає визначення які алелі в кількох локусах знаходяться на одній хромосомі, і ця інформація має велике значення в клінічній генетиці та генетиці популяцій. CliqueSNV розроблено, як новий підхід до реконструкції гаплотипів, спеціальний для роботи з низькочастотними варіантами.

CliqueSNV – використовує графовий підхід: спочатку ідентифікує всі варіанти одиночних нуклеотидних поліморфізмів (SNV) у вирівняних послідовностях, а потім групує їх у зв'язані, заборонені або некласифіковані пари (див. Рисунок 2.3). Зв'язані SNV зустрічаються достатньо часто, щоб припустити, що вони належать до одного гаплотипу. Заборонені SNV зустрічаються дуже рідко, що свідчить про те, що вони не можуть бути частиною одного гаплотипу. Некласифіковані SNV - це ті, які не можна надійно віднести до зв'язаних або заборонених.

Потім CliqueSNV будує граф з набором вузлів, які представляють SNV, і набором ребер, які з'єднують зв'язані пари SNV. Ідеально, SNV кожного правильно реконструйованого гаплотипу утворюють повний граф. Повний граф - це набір вузлів, в якому будь-які два вузли з'єднані ребром.[5]

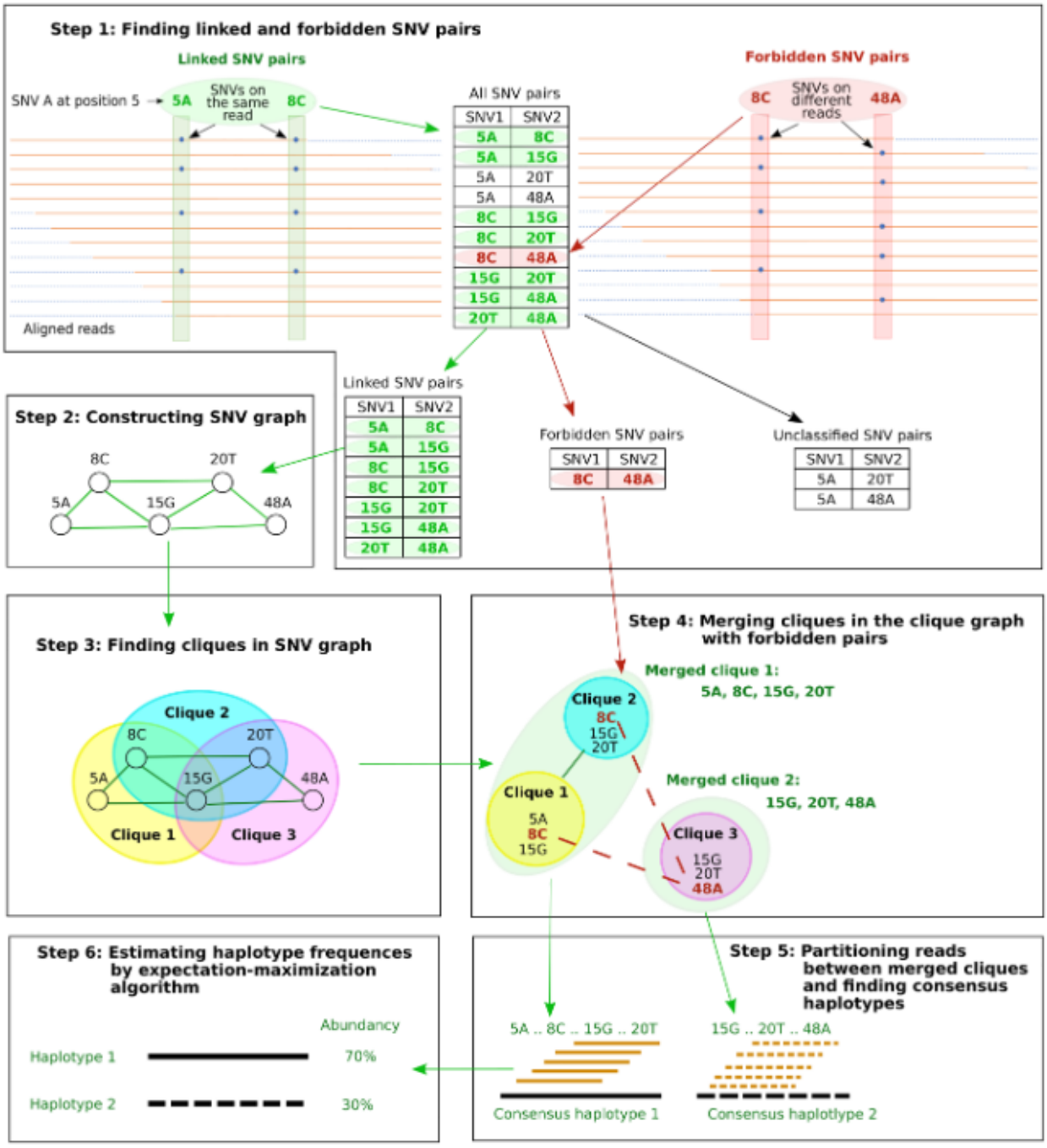



*Рисунок 2.3* Схема-алгоритм роботи CliqueSNV. [5]

### 2.2.2 PredictHaplo

PredictHaplo - це обчислювальний інструмент, що використовується в геноміці для реконструкції гаплотипів з послідовності даних.

PredictHaplo використовує ймовірнісну модель, що означає, що він використовує статистику для прогнозування найбільш ймовірного гаплотипу, враховуючи дані послідовності. Він використовує для цього приховану модель Маркова (HMM).

Приховані моделі Маркова - це статистичні моделі, де система, що моделюється, вважається процесом Маркова з невиявленими (прихованими) станами. У цьому випадку "прихована" частина відноситься до реальних послідовностей гаплотипів, а "спостережувана" частина відноситься до відомих даних генетичної послідовності.

PredictHaplo моделює еволюцію популяції гаплотипів вздовж окремої послідовності, використовуючи модель коалесценції з рекомбінацією. Він використовує цю модель для присвоєння постеріорних ймовірностей можливим гаплотипам, з урахуванням генотипу особини та карти рекомбінації.

HMM в PredictHaplo є наближенням до моделі коалесценції з рекомбінацією. У популяційній генетиці теорія коалесценції - це ретроспективна модель для відбудови генеалогії алелей. Ця модель припускає, що якщо ми простежимо за лінією двох алелей, вибраних з популяції, вони зійдуться у спільного предка. Модель коалесценції з рекомбінацією додатково дозволяє розривати та рекомбіновати хромосоми під час спадковості, що може викликати зміну генеалогії вздовж послідовності.

Рекомбінація відноситься до процесу, коли генетичний матеріал змішується під час статевого розмноження. Він важливий у формуванні генетичного різноманіття, оскільки дозволяє різні комбінації генів у потомстві. Карта рекомбінації - це карта, яка показує швидкість рекомбінації в різних позиціях вздовж хромосом. Ця інформація може бути корисною для визначення нерівноваги зв'язків (невипадкової асоціації алелей в різних локусах) та для генетичного картування.

Отже, теорія за PredictHaplo ґрунтується на ймовірнісній моделі, яка використовує приховану модель Маркова, що наближає модель коалесценції з рекомбінацією. Він присвоює постеріорні ймовірності можливим гаплотипам на основі генотипу особини та карти рекомбінації.[6]

### 2.2.3 aBayesQR

aBayesQR є інструментом Байєсівської Квантильної Регресії, що використовується в геноміці та біоінформатиці для картування Квантитативних Локусів Ознак (QTL). QTL - це ділянки ДНК, що корелюють з варіацією фенотипу або спостережуваної ознаки. Виявлення цих локацій може допомогти дослідникам зрозуміти генетичну основу варіації ознак, що є критично важливим у таких галузях, як медичні дослідження та сільське господарство.

Спершу зрозуміємо основні пов’язані концепції.

1. Квантильна регресія: Квантильна регресія (QR) - це статистична техніка, призначена для оцінки та роблення висновків про умовні квантильні функції. Так само, як класичні методи лінійної регресії, що базуються на мінімізації

сум квадратів залишків, дозволяють оцінювати моделі для умовних середніх функцій, методи квантильної регресії пропонують механізм для оцінювання моделей для умовної медіанної функції, а також повного спектра інших умовних квантильних функцій.

2. Байєсівська квантильна регресія: Байєсівська квантильна регресія (BQR) - це підхід, який вводить байєсівський висновок у квантильну регресію. Байєсівські методи відомі властивістю інтегрувати попередні знання та працювати із неоднозначними параметрами, що особливо цінне, коли набір даних маленький або система є складною.

У BQR, замість безпосередньої мінімізації функції втрат, як у класичному QR, використовуються байєсівські методи для оцінки апостеріорного розподілу параметрів. Це надає повну ймовірнісну модель, яка дозволяє проводити більш складну кількісну оцінку неоднозначності та моделювання прогнозування. У BQR обчислення, як правило, здійснюється за допомогою методів Марківських Ланцюгів Монте Карло (MCMC).

3. aBayesQR**:** Інструмент aBayesQR - це конкретне застосування Байєсівської Квантильної Регресії для генетичного картування. У геноміці дослідники часто хочуть зрозуміти взаємозв'язок між набором генетичних варіантів (наприклад, SNPs) і кількісною ознакою (наприклад, зріст). Традиційні методи, такі як дослідження Gene-Wide ASsociation (GWAS), зазвичай фокусуються на середньому ефекті, але вони можуть пропустити важливі ефекти в інших частинах розподілу.

aBayesQR долає це обмеження застосуванням Байєсівської Квантильної Регресії до проблеми, надаючи більш детальне розуміння генотип-фенотип мапи. Він застосовує байєсівський відбір змінних для автоматичного вибору відповідних SNPs, і використовує спеціальний клас апріорі (асиметричні Лапласівські апріорі) для більш ефективної оцінки квантильної регресії.

aBayesQR використовує методи MCMC для вибірки з апостеріорного розподілу, які потім можуть бути використані для висновків. Він також застосовує деякі передові обчислювальні методи для більш ефективного вибірки, такі як паралельне згладжування.

Підсумовуючи, aBayesQR - це потужний інструмент, який застосовує Байєсівську Квантильну Регресію до проблеми генетичного картування. Він надає більш детальне розуміння генотип-фенотип мапи і здатний автоматично вибирати відповідні SNPs, що робить його цінним інструментом у геноміці та біоінформатиці.[7]

## 2.3 Методи генного підрахунку

### 2.3.1 Kallisto

Розглянемо базові концепції генного підрахунку на прикладі kallisto.

Kallisto - це інструмент геноміки, розроблений для аналізу даних високопродуктивного секвенування РНК (RNA-Seq). Теоретичні основи Kallisto базуються на двох основних поняттях: псевдовирівнюванні рідів та кількісному визначенні транскриптів за допомогою алгоритму очікування-максимізація (EM). Ці концепції мають свої коріння в комп'ютерних науках, зокрема в зіставленні рядків, хешуванні та ймовірнісному моделюванні.

#### **2.3.1.1** Псевдовирівнювання

Центральна ідея псевдовирівнювання Kallisto полягає в використанні спрощеного представлення рідів і референтних геномів, що дозволяє заощадити час і обчислювальні ресурси порівняно з традиційними методами вирівнювання. Замість визначення точного місця ріда на референтному геномі або транскриптомі, Kallisto визначає набір транскриптів, з яких міг походити рід, на основі схожості k-мерів. K-мери - це послідовності довжиною k, що виведені з ріда.

Kallisto створює кольоровий граф де Бройна з транскриптома, де кожний транскрипт відповідає окремому кольору. Кожен рід потім може бути проаналізований як послідовність k-мерів та відображений як шлях в графі. Псевдовирівнювання ріда - це набір транскриптів (кольорів), що відповідають шляхам, які проходять k-мери ріда.

#### **2.3.1.2** Кількісне визначення транскриптів

Після отримання псевдовирівнювань, Kallisto використовує їх для оцінки кількості транскриптів. Це робиться за допомогою алгоритму очікування-максимізації (EM) для розв'язання системи лінійних рівнянь, де змінні представляють кількість транскриптів. Система рівнянь представляє взаємозв'язок між кількістю транскриптів та спостережуваною кількістю рідів.

Алгоритм EM ітеративно оцінює ці кількості транскриптів. Крок "очікування" оцінює внесок кожного транскрипта в спостережувані ріди на основі поточних оцінок кількості транскриптів. Крок "максимізація" потім оновлює оцінки кількості транскриптів на основі цих внесків. Алгоритм продовжується до того, як оцінки кількості транскриптів не збігаються.

#### **2.3.1.3** Теоретичні переваги

Використання Kallisto псевдовирівнювання та алгоритму EM має декілька теоретичних переваг. Етап псевдовирівнювання швидше та менше витрачає ресурси, ніж традиційне вирівнювання, при цьому забезпечуючи необхідну інформацію для кількісного визначення транскриптів. Використання алгоритму EM дозволяє Kallisto ефективно вирішувати проблему з рідами, які могли б походити від кількох транскриптів, загальною проблемою в даних RNA-Seq через альтернативний сплайсинг та інші процеси. Ймовірнісне моделювання також забезпечує природний спосіб включення невизначеності в оцінки кількості транскриптів, що призводить до більш надійних результатів.

Важливо згадати, що методологія Kallisto дозволяє інструменту проводити кількісне визначення безпосередньо на сирих даних рідів, оминаючи створення файлів вирівнювання, і значно знижуючи обчислювальне навантаження та вимоги до пам’яті. [8]

### 2.3.2 Gromstole

Gromstole - це набір скриптів Python та R, що розроблені для оцінки відносних частот різних ліній SARS-CoV-2, включаючи конкретні відомі небезпечні варіанти, за допомогою даних секвенування наступного покоління, отриманих зі зразків стічних вод. Варто відзначити, що у Gromstole відсутня офіційна математична документація або наукова стаття, що детально описує його методології. Натомість, його функціональні деталі в основному розміщені на GitHub, платформі, що часто використовується для управління вихідним кодом і контролю версій.

Gromstole складається з трьох основних скриптів: скрипта Python (minimap2.py), який діє як обгортка для програми маппінгу посилань minimap2; і двох скриптів R (estimate-freqs.R та make-barplots.R). Обгортка minimap2.py дозволяє швидко виводити дані виводу з minimap2, видобуваючи статистику покриття і частоту мутацій для кожного зразка. Скрипт estimate-freqs.R використовує квазібіноміальну регресію для оцінки частоти небезпечного варіанту, з частот мутацій, засновану на відповідному файлі з відповідними мутаціями у папці constelations. Скрипт make-barplots.R призначений для візуалізації цих оцінок частот варіантів в наборі зразків у вигляді стовпчастої діаграми. Втім, користь цих компонентів та їх взаємодія в межах фреймворку Gromstole залишається недостатньо задокументованою в офіційній академічній літературі.

# 3 Розділ 3. Результати досліджень

## 3.1 Створення еталонних наборів даних

Для отримання референсних геномів вірусу було використано дані з бази даних Global Initiative on Sharing Avian Influenza Data (GISAID), яка є публічною базою даних, що містить повні геномні послідовності вірусів, включаючи SARS-CoV-2. Доступ до бази даних був здійснений через веб-сайт GISAID, а дані були завантажені у форматі FASTA, тип файлу, який дозволяє здійснювати швидкий пошук за подібністю.[9]

Для генерації наборів даних було використано SWAMPy.

Згенеровані датасети:

**1) Simulated data set - SD2 (без помилок секвенування)**

Дані симульовані з Art

11 ліній, повний геном, 42 зразки, парні ріди, довжина ріду - 150 бп, 1000X (~1093023 рідів у кожному зразку):

**2) Simulated data set - SD3 (без помилок секвенування)**

Дані симульовані з Art

12 ліній, повний геном, 36 зразків, парні ріди, довжина ріду - 150 бп, 1000X (~1093023 рідів у кожному зразку):

**3) Simulated data set - SD4 (з помилками секвенування)**

Дані симульовані з Art

11 ліній, повний геном, 42 зразки, парні ріди, довжина ріду - 150 бп, 1000X (~1186200 рідів у кожному зразку)

За допомогою інструменту симуляції рідів SimSeq, було зганеровано кілька датасетів які використовувалися для аналізу інструментів реконструкції гаплотипів:

**1) Simulated data set - spike_mix5_50000 (з помилками секвенування)**

Дані симульовані з SimSeq

5 ліній, амплікон спайку, 1 зразок, парні ріди, довжина ріду - 150 бп, ~400X (50000 рідів у кожному зразку):

**2) Simulated data set - fullen_mix5_50000 (з помилками секвенування)**

Дані симульовані з SimSeq

5 ліній, повний геном, 1 зразок, парні ріди, довжина ріду - 150 бп, ~100X (50000 рідів у кожному зразку):

## 3.2 Порівняльний аналіз

### 3.2.1 Методи реконструкції гаплотипів

Запропоновано наступний пайплайн Рисунок 3.1.

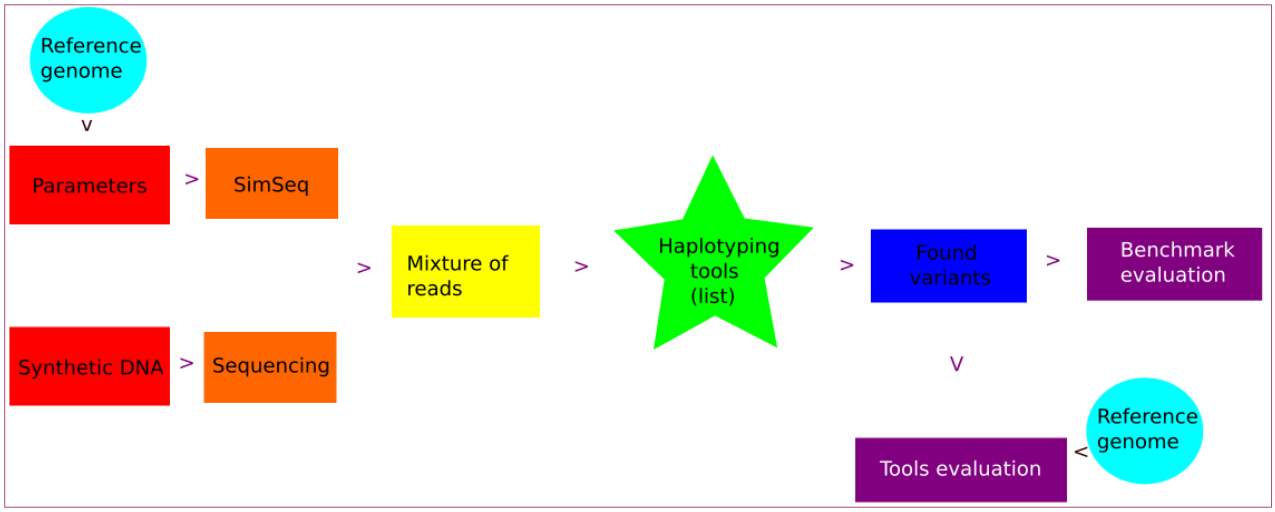



*Рисунок 3.2* Запропонований пайплайн зображає послідовність кроків від отримання наборів даних до розрахунку та аналізу результатів.

Для оцінки CliqueSNV, aBayesQR та PredictHaplo було використано spike_mix5_50000 та fullen_mix5_50000 набори даних.

На початковому етапі дані, за допомогою samtools bwa[10], були вирівняні по відношенню до загального референтного геному Sars-Cov-2 знайденого на GISAID. Було використано програму Tablet для візуального контролю вирівнювання (Рисунок 3.2).

За допомогою ClustalW було проведено мульти-послідовнісне вирівнювання для оцінки і підвищення якості знаходження правильних реконструйованих варіантів.

*Рисунок 3.3* Результат вирівнювання *spike_mix5_50000.*

В якості функції помилки ми беремо EMD (Earth Mover's Distance), в даному випадку спрощену до Hamming Distance, що визначається як сума помилкових пар основ у реконструйованій послідовності. В якості метрик ефективності методів використовуються влучність та повнота.

Влучність визначається, як

$$\text{Влучність} = \frac{tp}{tp + fp}$$

Повнота визначається, як

$$\text{Повнота} = \frac{tp}{tp + fn}$$

На Рисунку 6.3 зображені результати роботи порівнюваних інструментів.

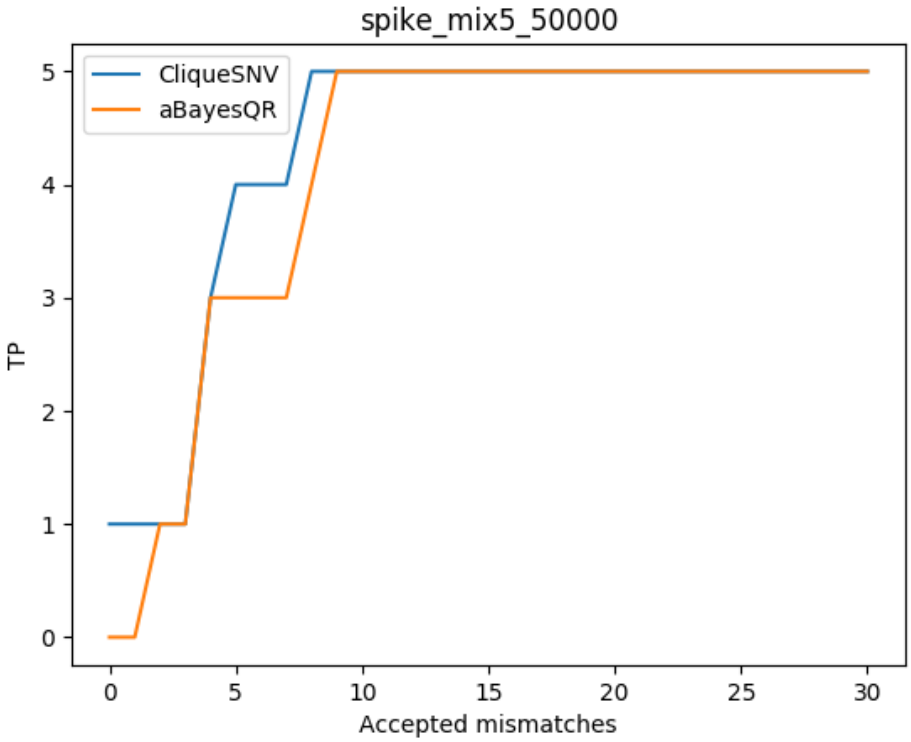
| | CliqueSNV | | | | PredictHaplo | | | | aBayesQR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Variants detected | EMD | Precision | Recall | Variants detected | EMD | Precision | Recall | Variants detected | EMD |
| Spike Mixture | 1/13 | 1/5 | 13 | 17 | 0 | 0 | 0 | - | 0 | 0 | 7 | 16 |
| Full genome Mixtures | 0 | 0 | 1 | 1497 | 0 | 0 | 5 | 2154 | 0 | 0 | 3 | 120 |

*Рисунок 3.4* Влучність (Precision) і повнота (Recall) оцінені для кожного із трьох інструментів на двох наборах даних.

Отже, PredictHaplo зависло на підрахунку для спайк датасету. На повногеномному датасеті значення EMD для всіх інструментів дуже великі.

На Рисунках 3.4 зображена залежність кількості знайдених дійсно позитивних і хибно позитивних гаплотипів від порогу EMD. На Рисунку 6.3 цей поріг – 0.

A




B

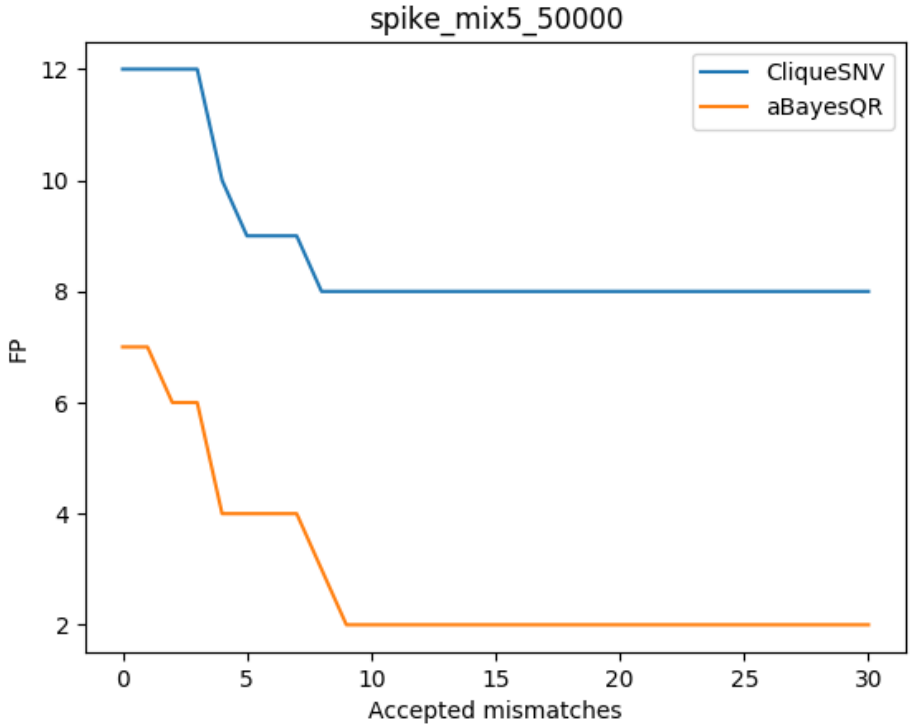



*Рисунок 3.5* Графіки показують динаміку зміни кількостей (A) дійсно позитивних і (B) хибно позитивних розпізнавань ліній по мірі збільшення числа допустимих невідповідностей між референтними і реконструйованими геномами.

### 3.2.2 Методи генного підрахунку

Для аналізу методу генного підрахунку gromstole було використано набори даних SD2 та SD4. На Рисунку 3.5 показана таблиця з дійсними пропорціями варіантів у зразках SD2.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | n | Alpha | Beta | Delta | Epsilon | Eta | Gamma | Iota | Kappa | Lambda | Theta | Zeta | Epsilon1 | |
| 2 | 1 | 0.86 | 17.85 | 0.12 | 98 | 0.06 | 7.15 | 23.82 | 0 | 11.95 | 0.3 | 0.37 | 90.85 | |
| 3 | 2 | 3.03 | 17.76 | 0.43 | 92.8 | 0.34 | 7.05 | 24.7 | 13.35 | 12.41 | 1.16 | 2.03 | 86.63 | |
| 4 | 3 | 7.66 | 17.47 | 0.66 | 88.74 | 0.38 | 7.08 | 24.18 | 12.78 | 12.75 | 1.98 | 1.38 | 83.34 | |
| 5 | 4 | 5.38 | 17.74 | 0.29 | 91.07 | 0.34 | 7.53 | 24.15 | 12.86 | 12.72 | 1.6 | 1.75 | 85.03 | |
| 6 | 5 | 5.08 | 17.66 | 0.54 | 91.86 | 0.42 | 6.98 | 24.36 | 13.55 | 12.71 | 1.39 | 1.17 | 85.9 | |
| 7 | 6 | 9.18 | 17.18 | 0.11 | 86.34 | 0.6 | 7.18 | 24.39 | 12.35 | 13.18 | 2.46 | 2.26 | 81.56 | |
| 8 | 7 | 12.07 | 17.38 | 0.03 | 82.29 | 0.74 | 7 | 25.29 | 11.03 | 13.74 | 3.1 | 2.24 | 78.35 | |
| 9 | 8 | 12.74 | 17.26 | 0.03 | 82.12 | 0.7 | 6.93 | 25.24 | 11.32 | 13.58 | 3.17 | 1.67 | 78.01 | |
| 10 | 9 | 19.38 | 17.14 | 0.07 | 74.17 | 0.79 | 7.32 | 25.37 | 9.63 | 14.55 | 4.94 | 2.88 | 71.79 | |
| 11 | 10 | 26.45 | 16.41 | 0.02 | 67.16 | 0.81 | 7.29 | 25.24 | 8.83 | 15.49 | 6.32 | 2.73 | 65.98 | |
| 12 | 11 | 36.76 | 15.63 | 0.08 | 55.57 | 1.24 | 7.5 | 24.93 | 6.69 | 16.89 | 8.42 | 2.93 | 56.95 | |
| 13 | 11 | 36.76 | 15.63 | 0.08 | 55.57 | 1.24 | 7.5 | 24.93 | 6.69 | 16.89 | 8.42 | 2.93 | 56.95 | |
| 14 | 13 | 49.91 | 14.71 | 0.06 | 42.09 | 1.35 | 8.25 | 24.76 | 4.46 | 18.23 | 11.34 | 2.91 | 45.93 | |
| 15 | 14 | 55.69 | 14.06 | 0.07 | 35.95 | 1.36 | 8.68 | 24.36 | 3.43 | 19.03 | 12.55 | 2.77 | 41.01 | |
| 16 | 15 | 59.65 | 14.03 | 0.06 | 30.69 | 1.6 | 9.82 | 24.8 | 2.32 | 19.79 | 13.94 | 3.2 | 36.92 | |
| 17 | 15 | 59.65 | 14.03 | 0.06 | 30.69 | 1.6 | 9.82 | 24.8 | 2.32 | 19.79 | 13.94 | 3.2 | 36.92 | |
| 18 | 17 | 67.31 | 12.87 | 0.29 | 23.11 | 1.54 | 12.05 | 22.39 | 1.53 | 20.68 | 15.78 | 2.92 | 30.59 | |
| 19 | 18 | 67.85 | 12.89 | 0.59 | 21.57 | 1.59 | 13.23 | 22.64 | 1.42 | 20.89 | 15.98 | 3.13 | 29.35 | |
| 20 | 19 | 68.85 | 12.48 | 1.2 | 20.26 | 1.62 | 14.03 | 21.99 | 1.63 | 21.53 | 16.37 | 2.96 | 28.4 | |
| 21 | 20 | 69.74 | 12.3 | 1.56 | 19.32 | 1.61 | 15.45 | 21.32 | 1.68 | 21.74 | 17.01 | 3.01 | 27.5 | |
| 22 | 21 | 69.93 | 12.22 | 2.24 | 18.49 | 1.63 | 16.2 | 20.64 | 2.05 | 21.52 | 17.18 | 2.97 | 26.85 | |
| 23 | 22 | 69 | 12.06 | 3.32 | 17.99 | 1.71 | 17.38 | 20.53 | 2.79 | 21.81 | 17.02 | 3.16 | 26.49 | |
| 24 | 23 | 67.57 | 11.77 | 5.25 | 17.16 | 1.84 | 18.12 | 20.01 | 4.02 | 21.86 | 17.03 | 3.19 | 25.75 | |
| 25 | 24 | 63.83 | 10.99 | 10.01 | 17.38 | 2.17 | 18.63 | 19.87 | 7.39 | 22.28 | 16.55 | 3.18 | 25.96 | |
| 26 | 25 | 55.44 | 9.96 | 20.18 | 17.24 | 2.83 | 18.29 | 19.73 | 14.79 | 22.41 | 14.91 | 3.18 | 25.82 | |
| 27 | 26 | 48.11 | 8.2 | 32.43 | 17.63 | 3.32 | 16.05 | 19.08 | 23.24 | 22.57 | 12.65 | 2.54 | 26.51 | |
| 28 | 27 | 38.02 | 6.27 | 46.18 | 18.4 | 3.95 | 13.65 | 18.79 | 33.22 | 23.18 | 10.16 | 2.08 | 27.32 | |
| 29 | 28 | 23.33 | 4.24 | 64.82 | 18.98 | 5.04 | 10.62 | 18.53 | 46.44 | 23.14 | 6.8 | 1.67 | 27.85 | |
| 30 | 29 | 14.7 | 2.71 | 77.45 | 20.04 | 5.69 | 7.05 | 19 | 55.38 | 23.37 | 4.35 | 1.11 | 29.03 | |
| 31 | 30 | 7.55 | 1.42 | 87.77 | 20.52 | 6.13 | 4.05 | 18.83 | 62.74 | 23.25 | 2.35 | 0.58 | 29.42 | |
| 32 | 31 | 3.9 | 0.79 | 93.19 | 0 | 6.48 | 2.39 | 19 | 66.41 | 23.53 | 1.26 | 0.38 | 0 | |
| 33 | 32 | 1.98 | 0.44 | 96.28 | 0 | 6.4 | 1.45 | 18.88 | 85.39 | 23.58 | 0.66 | 0.21 | 0 | |
| 34 | 33 | 1.07 | 0.23 | 97.73 | 0 | 6.58 | 0.8 | 18.78 | 90.28 | 23.39 | 0.38 | 0.11 | 0 | |
| 35 | 34 | 0.49 | 0.2 | 98.28 | 0 | 6.75 | 0.55 | 19.04 | 89.23 | 23.42 | 0.24 | 0.12 | 0 | |
| 36 | 35 | 0.25 | 0.08 | 99.08 | 0 | 6.75 | 0.27 | 0 | 88.7 | 23.41 | 0.11 | 0.04 | 0 | |
| 37 | 36 | 0.14 | 0.06 | 99.07 | 21.29 | 0 | 0.17 | 0 | 90.91 | 0 | 0.06 | 0.03 | 30.21 | |
| 38 | 37 | 0.08 | 0.04 | 99.35 | 0 | 6.53 | 0.12 | 0 | 88.14 | 0 | 0.06 | 0.03 | 0 | |
| 39 | 38 | 0.04 | 0.02 | 99.43 | 0 | 0 | 0.06 | 0 | 89.94 | 0 | 0.02 | 0.01 | 0 | |
| 40 | 39 | 0.02 | 0 | 99.53 | 0 | 0 | 0.03 | 0 | 88.15 | 0 | 0.01 | 0 | 0 | |
| 41 | 40 | 0.05 | 0.04 | 99.32 | 21.43 | 0 | 0.05 | 0 | 88.44 | 0 | 0.02 | 0 | 30.32 | |
| 42 | 41 | 0 | 0.01 | 99.51 | 0 | 0 | 0.01 | 0 | 87.75 | 0 | 0.01 | 0.01 | 0 | |
| 43 | 42 | 0.01 | 0.01 | 99.48 | 0 | 0 | 0.02 | 0 | 90.05 | 0 | 0.01 | 0.01 | 0 | |
| 44 | | | | | | | | | | | | | | |

*Рисунок 3.6* Дані датасету SD2

Інструмент погано підтримувався, доступний лише «демо режим»: мутації тільки для cB.1.617.2 варіанту. Інші профілі мутацій Sars-Cov-2, що за замовчуванням до ступні у gromstole були виправлені до працюючого стану: нотація позиції мутацій змінена, прибрані мутації вставки і видалення.

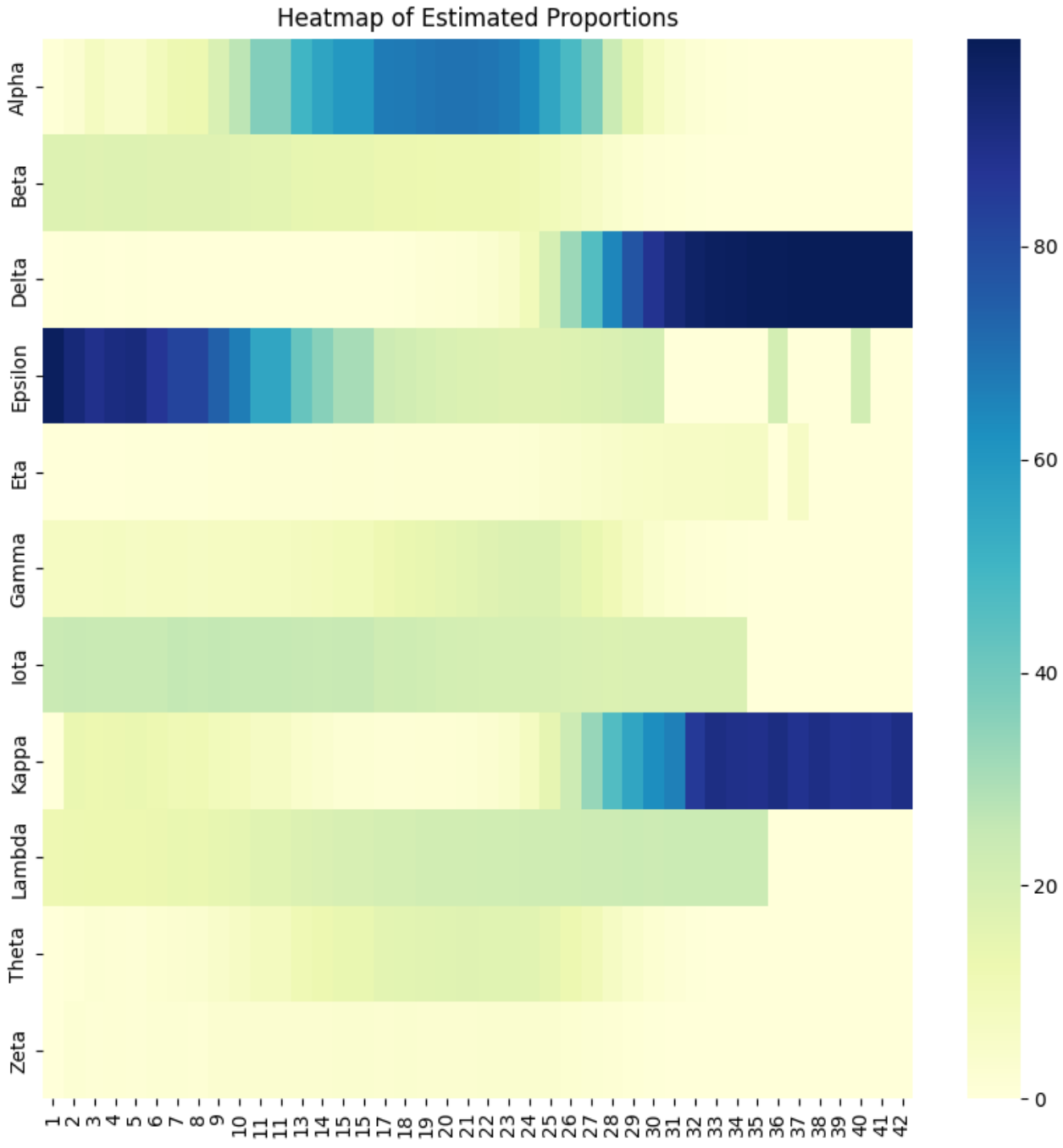



*Рисунок 3.7* Теплова карта оцінок концентрації для усіх зразків та ліній.

На Рисунку 3.6 і Рисунку 3.7 візуалізовано результати роботи gromstole.

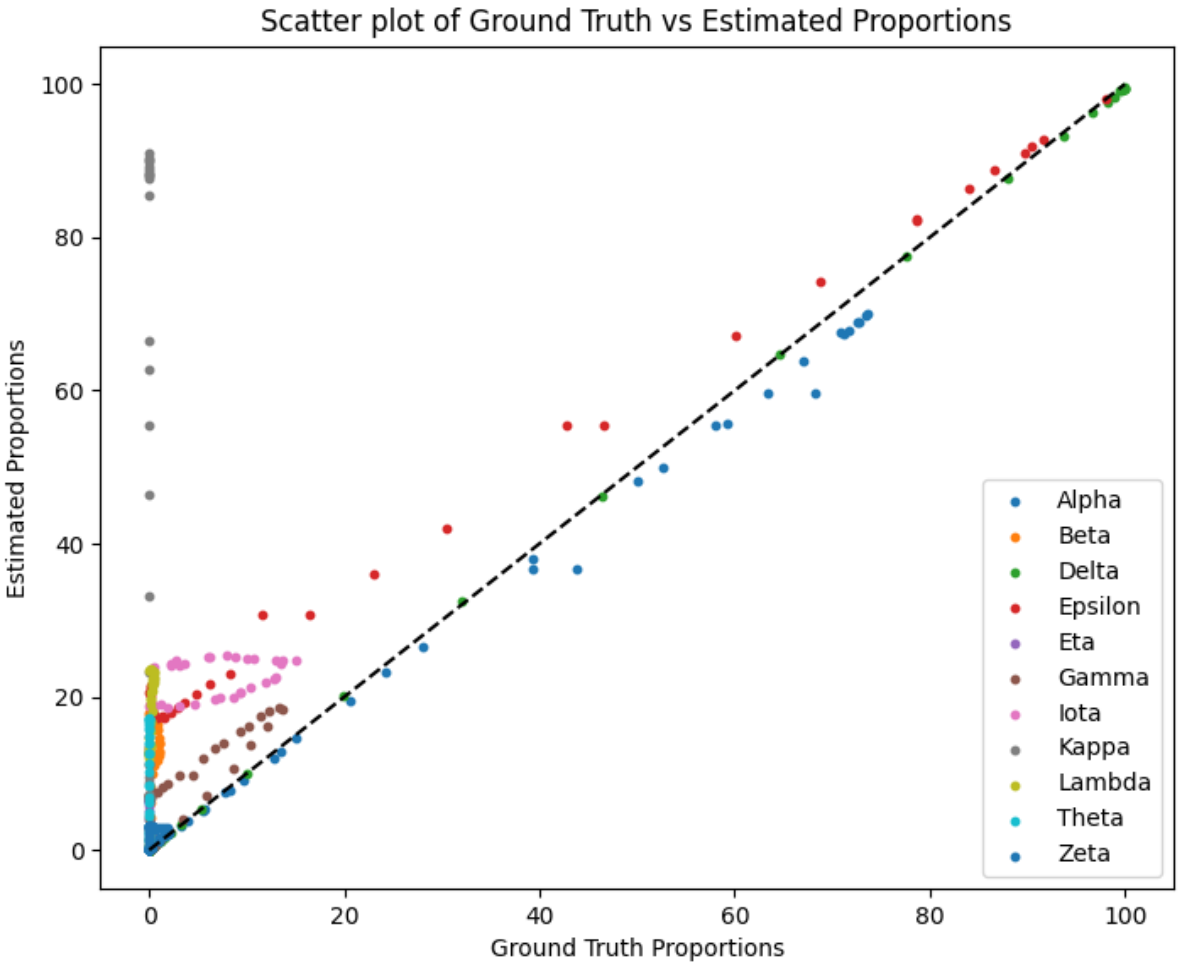



*Рисунок 3.8* Графік, що відображає усі 42 зразки, з реальними значеннями концентрації усіх ліній по горизонталі і розрахованими їх значеннями по вертикалі.

## 3.3 Майбутні напрямки

Для закінчення роботи над CliqueSNV, PredictHaplo та aBayesQR потрібно оцінити їх роботу на SD2, SD4 датасетах. Це дозволить здійснити статистичний аналіз результатів, за рахунок великих об`ємів цих наборів даних.

У майбутньому можна проаналізувати методи реконструкції які було систематизовано у статті [11], список на *Рисунок 3.8*.

Table 2. Haplotype calling software tools for viral NGS data

| Haplotyping tools | Year | System | *De novo*/Ref based | Pair-end reads | Sequencing error handling | Haplotype assembly method | Haplotype frequency estimation method | Output sequences |
|---|---|---|---|---|---|---|---|---|
| Shorah [82] | 2011 | Linux | Ref | + | Probabilistic clustering | Minimal path cover | EM | Full haplotypes |
| ViSpA [60] | 2011 | Linux | Ref | – | Binomial model | Max-bandwidth path | EM | Full haplotypes |
| QColors [86] | 2012 | — | *De novo* | – | — | Overlap graph + Conflict graph | — | Full haplotypes |
| QuRe [87] | 2012 | Java | Ref | + | Poison model | Multinomial distribution matching | Read coverage | Full haplotypes |
| bioa [85] | 2012 | Linux | Ref | – | k-mer-based error correction | Maximum Bandwidth Path | Fork balancing | Full haplotypes |
| Vicuna [63] | 2012 | Linux | *De novo* | + | Read count | — | — | Consensus + contigs |
| QuasiRecomb [95] | 2013 | Linux | Ref | + | Hidden Markov model | Hidden Markov model | Hidden Markov model | Full haplotypes |
| Vira (AmpMCF) [88] | 2013 | Linux | Ref | – | — | Multicommodity flows | Normalized flow size | Full haplotypes |
| ShotMCF [88] | 2013 | JAVA | Ref | – | Binomial model | Max-bandwidth path + Multicommodity flows | EM + normalized flow size | Full haplotypes |
| BAsE-Seq [61] | 2014 | – | Ref | + | Poisson binomial distribution model | Clustering of reads by SNVs | Read coverage | Full haplotypes |
| VGA [90] | 2014 | Linux | Ref | + | Requires high-fidelity sequencing protocol | Min-graph coloring | EM | Full haplotypes |
| HaploClique [89] | 2014 | Linux | Ref | + | — | Max-clique enumeration | Normalized read count | Full haplotypes |
| PredictHaplo [96] | 2014 | Linux | Ref | + | Dirichlet Process Mixture Model | Dirichlet Process Mixture Model | Dirichlet Process Mixture Model | Full haplotypes |
| IVA [64] | 2015 | Linux | *De novo* | – | Read count | — | — | Contigs |
| MLEHaplo [98] | 2015 | Linux | *De novo* | + | — | Maximum likelihood | — | Full haplotypes |
| ViQuaS [91] | 2015 | Linux | Ref | + | Chimeric error correction | Multinomial distribution matching | Read count | Full haplotypes |
| SAVAGE [65] | 2017 | Linux | *De novo* | + | Overlap fuzzy matching error correction | Enumerating cliques in overlap graph | EM | Contigs |
| aBayesQR [100] | 2017 | Linux | Ref | + | Cluster coverage by reads | Bayesian inference | Bayesian inference | Full haplotypes |
| RegressHaplo [97] | 2017 | R | Ref | + | — | Penalized regression | Penalized regression | Full haplotypes |
| 2SNV [99] | 2017 | Java | Ref | – | Linkage of SNV pairs | Hierarchical clustering of reads by SNVs | EM | Full haplotypes |
| PEHaplo [92] | 2018 | Linux | *De novo* | + | Overlap error correction | Path finding in overlap graph | — | Contigs |
| Shiver [66] | 2018 | Linux | *De novo* + ref | + | BLAST database match | — | — | Consensus |
| CliqueSNV [76] | 2018 | JAVA | Ref | + | Linkage of SNV pairs | Clique enumeration and merging | EM | Full haplotypes |

*Рисунок 3.8 Список інструментів для реконструкції гаплотипів* [11]

Для порівняльного аналізу розглянутих інструментів, а також для можливості додавання інструментів і даних іншими науковцями розглядається настройка пайплайну в Omnibenchmark (див. Рисунок 3.9).

*A* *B*

```
## modules
from omnibenchmark.utils.build_omni_object import get_omni_object_from_yaml

## Load object
omni_obj = get_omni_object_from_yaml('src/config.yaml')
```

```
---
data:
    name: "module-name"
    title: "A new module"
    description: "A new module for omnibenchmark, e.g., a dataset, method, metric,..."
    keywords: ["module-type-key"]
script: "path/to/module_script"
outputs:
    template: "data/${name}/${name}_${out_name}.${out_end}"
    files:
        counts:
            end: "mtx.gz"
        data_info:
            end: "json"
        meta:
            end: "json"
benchmark_name: "an-omnibenchmark"
```

*C*

```
## create output dataset that stores all result/output files
omni_obj.create_dataset()

## Update inputs from other modules
omni_obj.update_obj()

## Run your script with all defined inputs and outputs.
## This also generates a workflow description (plan) and is tracked as activity.
omni_obj.run_renku()

## Link output files to output dataset
omni_obj.update_result_dataset()

## Save and commit to gitlab
renku_save()
```

*Рисунок 3.9* Приклади представлення (A) модулів, (B) конфігураційних файлів і (C)пайплайнів у Omnibenchmark.

Інший інструмент, що використовують для пришвидшення та додавання гнучкості у роботі над порівняльним аналізом, а саме інтенсивним запуском програм на комп`ютерних кластерах, - це Snakemake. Він дозволяє зберігати і запускати пайплайни у вигляді Python файлів. Так можна зменшити кількість баш скриптів проекту звівши все в Python пайплайни і json файли зі змінними. Всі параметри і додаткову обробку даних можна додавати прямо у скрипт Snakemake пайплайну, що є зручно.

# ВИСНОВКИ

1) Було згенеровано масивні штучні набори даних рідів зі зразками 11ти варіантів SARS-COV-2. Ці набори в майбутньому можна використовувати для визначення якості програм для знаходження варіантів і їх концентрацій.
2) Було оцінено роботу методів CliqueSNV, PredictHaplo, aBayesQR для задачі детекції варіантів вірусів на штучному наборі даних.

   Ці методи погано справляються з набором даних повного геному: реконструкції PredictHaplo й CliqueSNV мають велику кількість невідповідностей, aBayesQR – на порядок менше, але методи не видали жодного дійсно позитивного варіанту.

   Щодо набору даних області спайку, також жодного дійсно позитивного варіанту. Було досліджено збільшення допустимої кількості невідповідностей між реконструйованими і референтними геномами. Збільшення порогу на від 1-3 одиниці збільшило кількість дійсно позитивних варіантів. Збільшення порогу більш ніж на 10 одиниць перестало додавати дійсно позитивні, проте ще лишаються хибно негативні.

   Варто зазначити, що PredictHaplo не зміг завершити роботу над спайк даними. Проблеми з цим інструментом виникали і на інших ітераціях датасетів.
3) Було оцінено роботу інструменту gromstole. Із крапкового графіку можна зробити деякі висновки. Варіант Kappa не було розпізнано у зразках; концентрацію деяких варіантів було передбачено гірше ніж інших. Це може бути зв’язано із якістю і повнотою профілів мутацій, є можливість їх покращити. Загалом, метод відпрацював нормально.

# 4 Список використаних джерел


[1] S. Karthikeyan *et al.*, ‘Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission 1 2 3’, *Rebecca Fielding-Miller*, vol. 18, doi: 10.1101/2021.12.21.21268143.

[2] S. Benidt and D. Nettleton, ‘SimSeq: A Nonparametric Approach to Simulation of RNA-Sequence Datasets’. [Online]. Available: http://cran.rproject.org/

[3] W. Boulton, F. R. Fidan, N. De Maio, and N. Goldman, ‘SWAMPy: Simulating SARS-CoV-2 Wastewater Amplicon Metagenomes with Python’, doi: 10.1101/2022.12.10.519890.

[4] W. Huang, L. Li, J. R. Myers, and G. T. Marth, ‘ART: a next-generation sequencing read simulator’. [Online]. Available: http://bioinformatics.oxfordjournals.org/

[5] S. Knyazev *et al.*, ‘Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction’, *Nucleic Acids Res*, vol. 49, no. 17, pp. E102–E102, Sep. 2021, doi: 10.1093/nar/gkab576.

[6] S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth, ‘HIV haplotype inference using a propagating dirichlet process mixture model’, *IEEE/ACM Trans Comput Biol Bioinform*, vol. 11, no. 1, pp. 182–191, 2014, doi: 10.1109/TCBB.2013.145.

[7] S. Ahn and H. Vikalo, ‘ABayesQR: A Bayesian method for reconstruction of viral populations characterized by low diversity’, in *Journal of Computational Biology*, Mary Ann Liebert Inc., Jul. 2018, pp. 637–648. doi: 10.1089/cmb.2017.0249.

[8] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, ‘Near-optimal probabilistic RNA-seq quantification’, *Nat Biotechnol*, vol. 34, no. 5, pp. 525–527, May 2016, doi: 10.1038/nbt.3519.

[9] Y. Shu and J. McCauley, ‘GISAID: Global initiative on sharing all influenza data – from vision to reality’, *Eurosurveillance*, vol. 22, no. 13. European Centre for Disease Prevention and Control (ECDC), Mar. 30, 2017. doi: 10.2807/1560-7917.ES.2017.22.13.30494.

[10] P. Danecek *et al.*, ‘Twelve years of SAMtools and BCFtools’, *Gigascience*, vol. 10, no. 2, Feb. 2021, doi: 10.1093/gigascience/giab008.

[11] S. Knyazev, L. Hughes, P. Skums, and A. Zelikovsky, ‘Epidemiological data analysis of viral quasispecies in the next-generation sequencing era’, *Briefings in Bioinformatics*, vol. 22, no. 1. Oxford University Press, pp. 96–108, Jan. 01, 2021. doi: 10.1093/bib/bbaa101.