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

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

ПОРІВНЯННЯ ДВОХ ОБЧИСЛЮВАЛЬНИХ АЛГОРИТМІВ ДЛЯ ПІДВИЩЕННЯ ТОЧНОСТІ ПРОГНОЗУВАННЯ СПЕЦИФІЧНИХ ДЛЯ РАКУ НЕОЕПІТОПІВ, ОТРИМАНИХ ШЛЯХОМ АБЕРАНТНОГО АЛЬТЕРНАТИВНОГО СПЛАЙСИНГУ

Випускна кваліфікаційна робота магістра
студентки спеціальності 091 Біологія
ОП «Біоінформатика і структурна біологія»
Юрчикової Марії Олексіївни

Науковий керівник від кафедри
доцент кафедри молекулярної біотехнології
та біоінформатики
к.б.н., доц. О.Ю. Нипорко

Робота виконана у відділі біомедичної інформатики
Кафедри комп’ютерних наук (D-INFK) ETH Zürich
під керівництвом, проф., Gunnar Rätsch

Оцінка захисту роботи

_____

Київ − 2023 р.

АНОТАЦІЯ

Юрчикова М.О. Порівняння двох обчислювальних алгоритмів для підвищення точності прогнозування специфічних для раку неоепітопів, отриманих шляхом аберантного альтернативного сплайсингу. – Випускна кваліфікаційна робота магістра за спеціальністю 091 Біологія ОП «Біоінформатика і структурна біологія».

Метою цієї роботи було порівняння результатів двох обчислювальних алгоритмів, заснованому на сполученнях екзонів у варіантах сплайсингу (JP) і на основі графів сплайсингу (GP) для досягнення відтворюваності їх результатів і тим самим підвищення точності прогнозування пухлиноспецифічних неоепітопів, отриманих з альтернативного сплайсингу.

Тема роботи є надзвичайно актуальною, оскільки розвиток сучасних обчислювальних методів для прогнозування онкомаркерів, придатних для терапії, є багатообіцяючими та може потенційно принести велику користь у клінічних дослідженнях з розробки вакцин проти пухлин або біомаркерів, які також можуть бути використані для моніторингу пацієнтів після діагностики раку. Однак, щоб надати дослідникам надійні дані потрібна підверджена відтворюваність результатів біоінформатичних алгоритмів з можливістю повторення експериментів.

Були отримані цікаві результати, а саме:

1) Співпадіння значень експресій спільних послідовностей неоепітопів (9-мерів) для двох вихідних даних агоритмів і координат сполучення екзонів є доказом правильності підрахунку експресій на етапі вирівнювання з референсним геномом і їх нормалізації, а також визначення і екстракції координат, що свідчить про точність двох методів;
2) Кількості відфільтрованих пухлиноспецифічних, спільних для двох агоритмів 9-мерів демонструють різні ефекти фільтраційних параметрів, у випадках застосування строгих правил до порогового значення

експресії 9-мерів у когорті нормальних зразків тканин і в той же час «поблажливість» щодо вимог до експресії для включення 9-мерів у когорту ракових зразків впливає на кількість кандидатів у фідфільтрованій вибірці.

Підсумки порівняльного аналізу будуть використані у досягненні відтворюваності результатів двох методів і нададуть можливість продовження подальших in-silico досліджень з валідації експресії ідентифікованих неоепітопів за допомогою мас-спектрометрії та прогнозування афінності зв’язування з комплексом МНС.

Ключові слова: обчислювальний алгоритм, аберантний альтернативний сплайсинг, пухлиноспецифічні неоепітопи, відтворюваність результатів, in-silico дослідження з валідації

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

GP - Graph-based pipeline, алгоритм на основі графів сплайсингу

JP - Junction-based pipeline, алгоритм на основі сполучень екзонів у сплайс-варіантах

АК - амінокислоти

MHC I - головний комплекс гістосумісності класу I

HLA I - лейкоцитарний антиген людини I

TAA - пухлиноасоційовані антигени

TSA - пухлиноспецифічні антигени

TCGA - проєкт геноміки раку “Атлас геному раку”

GTEx - проєкт експресії генотипу та тканини

AC - альтернативний сплайсинг

BRCA - рак молочної залози

ICIs - інгібітори імунних контрольних точок

TAP - транспортер, пов'язаний із системою обробки антигену

EP - ендоплазматичний ретикулум

CD8+ - CD8+ Т-ефекторні клітини

TCR - Т-клітинні рецептори

CTLA-4 - цитотоксичний Т-лімфоцит-асоційований білок 4

PD-1 - білок запрограмованої клітинної смерті 1

SRSF1 - збагачений серином і аргініном фактор сплайсингу 1

SFRS10 - фактор сплайсингу, багатий на аргінін/серин 10

NMD - безглуздий розпад мРНК

MYC - протоонкоген MYC

SRA - архів читання послідовності

MS - мас-спектрометрія

ВСТУП

Імунотерапія здійснила революцію в галузі лікування раку, підвищивши виживаність пацієнтів із смертельними онкологічними захворюваннями. Одним з її основних перспективних напрямків є розробка персоналізованих неоантигенних вакцин. Неоантигени- це пухлиноспецифічні антигени, що виникають в результаті мутацій, прикладом яких є аберантний альтернативний сплайсинг. Вакцини, створені на основі неоантигенів мають кілька переваг. По-перше, вони здатні викликати дійсно пухлиноспецифічні Т-клітинні відповіді, тим самим не пошкоджуючи здорові тканини. По-друге, неоантигени ще називають епітопами de novo, що дає можливість обійти центральну толерантність Т-клітин до власних епітопів і таким чином знешкодити пухлину [1].

Сучасні обчислювальні алгоритми для з прогнозування придатних кандидатів для терапії неоантигенів отриманих зі специфічних для раку сплайс-варіантів м-РНК є багатообіцяючими. В їх основі лежить ідентифікація аберантних сполучень екзонів. Складність полягає у правильності ідентифікації справжніх позитивних сполучень, які є специфічними для раку, зводячи до мінімуму хибнопозитивні результати. Тому алгоритми враховують порівняння з подіями сплайсингу у здорових зразках пацієнтів. Однак більшість таких інструментів знаходяться у стадії розробки і передбачені ними неопітопи потребують валідації та перевірки імунногенності у лабораторних екпериментах in-vitro.

Таким чином метою роботи було досягти відтворюваності результатів двох обчислювальних алгоритмів, заснованому на сполученнях у сплайс-варіантах (JP) і на основі графів сплайсингу (GP) для підвищення точності прогнозування онкоспецифічних неоепітопів, отриманих шляхом аберантного альтернативного сплайсингу.

Відповідно були поставлені наступні задачі:

1) Проаналізувати результати двох алгоритмів і оцінити їх точність;

2) Дослідити вплив різних наборів фільтраційних параметрів на розмір даних вибірки потенційних пухлиноспецифічних кандидатів неоепітопів;
3) Визначити біологічну релевантність результатів пухлиноспецифічних кандидатів неоепітопів, які є спільно ідентифіковані обома методами.

РОЗДІЛ 1

1.1 Імунотерапія раку

Імунотерапія раку діє шляхом стимулювання іммуної системи організму для знешкодження аномальних клітини, отриманих із соматичних геномних мутацій [2]. Цей механізм ініціюється, коли CD8+ Т-ефекторні клітини ідентифікують аномальні білкові послідовності, які трансформуються в короткі епітопи та представлені на молекулі головного комплексу гістосумісності MHC класу I пухлинних клітин, також відомого як людський лейкоцитарний антиген (HLA) у людей (Рис. 1.1). Процес починається з деградації ендогенно синтезованих білків у пухлинній клітині протеасомами до більш коротких послідовностей 8-11 АК. Ці менші пептиди піддаються подальшому розщепленню пептидазами в цитозолі та ендоплазматичному ретикулумі та потрапляють в ЕР через транспортер, пов’язаний з комплексом обробки антигену (TAP). У ЕР пептиди зв’язуються зі змінною з MHC класу I. Комплекси пептид-MHC класу I потім транспортуються до плазматичної мембрани через комплекс Гольджі, де пептид може бути розпізнаний CD8+ цитотоксичними Т-клітинами [3]. Хоча деякі Т-клітини розпізнають антигени, спільні для нормальних і пухлинних клітин, Т-клітинні рецептори (TCR) зазвичай зв’язують неоантигени з вищою спорідненістю, і пухлини, які експресують більше неоантигенів, з більшою ймовірністю викликають імуноопосередковане усунення пухлини[4].

У міру прогресування пухлина набуває резистентності, проявами якої може бути зміна мікрооточення або видалення імуногенних клітин, які пригнічують шляхи імунної активації та посилюють імуносупресивні шляхи. На останній стадії розвитку пухлина набуває низької імуногенності, що спричинює неконтрольований ріст [5].

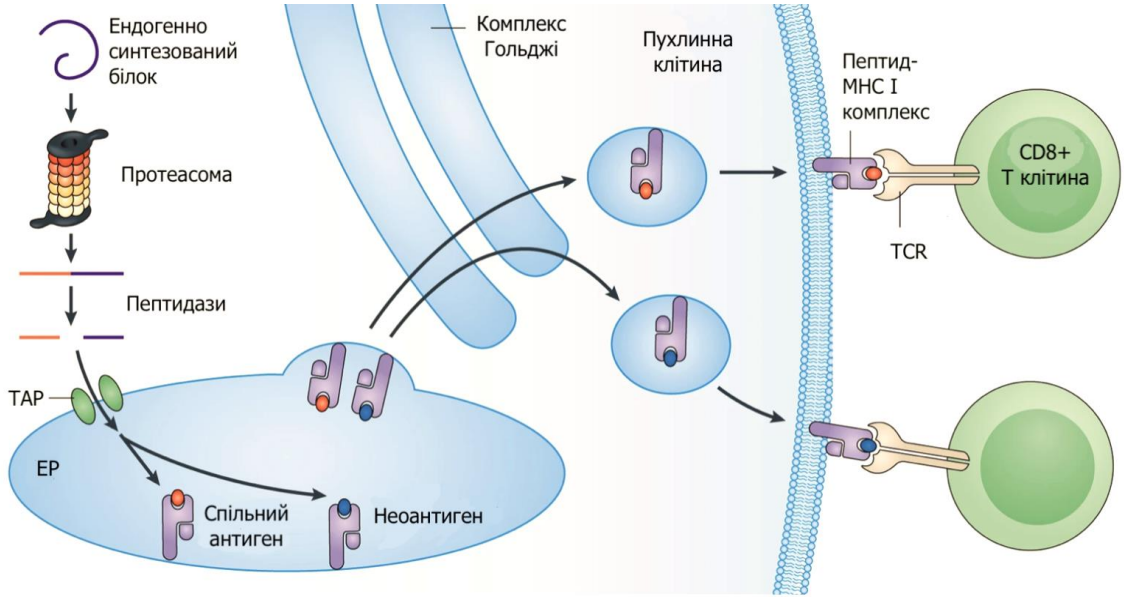

Рис. 1.1. Обробка та презентація неонтигену на комплексі MHC класу I.

Імунотерапія охоплює різні методи лікування, включаючи терапію інгібіторами імунних контрольних точок. Ця терапія передбачає введення препаратів, які блокують імунні контрольні точки, таких як антитіла CTLA-4 або PD-1, щоб зменшити імуносупресію. Однак поєднання антитіл CTLA-4 і PD-1 може підвищити ризик токсичності, і дослідження тривають для визначення предикторів ефективності та токсичності. Інший підхід полягає у введенні неоантигенів у сприятливе для імунітету мікрооточення, щоб налаштувати Т-клітини проти пухлини [6-7].

1.2 Категорії епітопів на поверхні пухлинної клітини

Можна поділити на такі типи як: 1) пухлиноасоційовані антигени (TAA) - це антигени, які експресуються як на пухлинних клітинах, так і на нормальних клітинах, вони можуть виникати внаслідок надмірної експресії нормальних білків або аберрантної експресії зазвичай мовчазних генів [7];2) пухлиноспецифічні антигени (TSA) – неоантигени (неоепітопи), які є унікальними для пухлинних клітин і не експресуються на нормальних клітинах. Вони можуть виникнути через соматичні мутації в ДНК пухлинної клітини і є ідеальними мішенями для імунотерапії раку, забезпечуючи найвищу протипухлинну імунну відповідь і найменший шанс спровокувати аутоімунну відповідь[7-8].

Термін неоепітоп використовується як синонім неоантигену, хоча під неоепітопом мають на увазі специфічні ділянки або послідовності неоантигену, які розпізнаються Т-клітинами або антитілами. Неоепітопи є справжніми мішенями імунної відповіді та відповідають за розпізнавання та усунення пухлинних клітин. Т-клітини можна налаштувати проти пухлини шляхом введення неоантигенів, які будуть націлені на нове мікрооточення, що сприятиме імунітету[9-10]. Тому точна ідентифікація потенційних цільових неоепітопів є важливою для ефективної імунотерапевтичної відповіді. Щоб бути розпізнаними невласними Т-клітинами, вони повинні бути присутніми виключно у пухлинних клітинах і мати сильну афінність зв’язування з комплексом MHC I класу[11].

1.3 Аберантний сплайсинг мРНК

Альтернативний сплайсинг - це процес, за допомогою якого кілька інформаційних РНК (мРНК) генеруються з однієї пре-мРНК, що призводить до функціонально відмінних білкових продуктів, які можуть виконувати різні або навіть протилежні ролі[12]. Доступність даних із таких джерел, як TCGA та проект GTEx, сприяла систематичним дослідженням кореляції явища АС з розвитком ракових захворювань. Понад 95% генів, транскрибованих у людини, піддаються альтернативному сплайсингу РНК, що призводить до

збільшення різноманітності протеому [13]. Процес експресії генів, відомий як аберантний альтернативний сплайсинг, порушує нормальний сплайсинг пре-мРНК, що призводить до генерації аномальних транскриптів мРНК [13-14].

1.4 Причини порушення регуляції сплайсингу при раку

Одніє з них є аберантна експресія факторів сплайсингу. Пухлинні клітини можуть мати змінені рівні експресії факторів сплайсингу, що може призвести до змін у моделях сплайсингу. Наприклад, надмірна експресія фактора сплайсингу SRSF1 була пов’язана зі збільшенням пропуску екзонів при раку [15].

Мутації у факторах сплайсингу або сайтах сплайсингу можуть порушити нормальні моделі сплайсингу та призвести до виробництва аберантних транскриптів. Наприклад, мутації у факторі сплайсингу SF3B1 були виявлені в кількох типах раку та пов’язані зі зміненими моделями сплайсингу[16].

Епігенетичні модифікації, такі як метилювання ДНК і модифікації гістонів, можуть впливати на доступність факторів сплайсингу до їх цільових сайтів і змінювати схеми сплайсингу. Наприклад, гіперметилювання ДНК промоторної області фактора сплайсингу SFRS10 було пов’язано зі зміненими моделями сплайсингу при раку молочної залози[17].

Альтернативний сплайсинг онкогенів або супресорів пухлин може генерувати ізоформи онкогенів або супресорів пухлин, які мають змінені функції порівняно з канонічною ізоформою. Наприклад, альтернативний сплайсинг онкогену MYC може генерувати коротшу ізоформу, яка є більш онкогенною[18].

Порушення регуляції шляхів процесингу РНК, наприклад безглуздий розпад (NMD) і редагування РНК, також може впливати на схеми сплайсингу при раку. NMD може призвести до деградації транскриптів із передчасними

стоп-кодонами, що може вплинути на моделі сплайсингу шляхом зміни рівнів факторів сплайсингу[19].

1.5 Визначення «нормального» і «пухлиноспецифічного» сплайсингу

У контексті ідентифікації неоепітопів постає питання що визначати як пухлиноспецифічним або нормальним явищем альтернативного сплайсингу. Оскільки є ризик помилкової ідентифікації значної кількісті пухлинноспесифічних сплайс-варіантів якщо порівнювати сполучення екзонів у когорті пухлинних зразків лише з даними нормальних зразків відповідного типу тканини. До когорти нормальних зразків бажано включати зразки всіх типів нормальних тканин людини які можна отримати з GTEx та з архіву прочитаної послідовності SRA[20-21].

1.6 Прогнозування неоантигенів in silico. Ідентифікація імуногенних неоантигенів

Ідентифікація імуногенних неоантигенів має вирішальне значення для розробки персоналізованої імунотерапії раку. Було розроблено методи in silico для прогнозування неоантигенів на основі генетичних мутацій пухлини та типу лейкоцитарного антигену людини (HLA) пацієнта[22]. Однак ці методи мають обмеження, такі як хибнопозитивні прогнози та труднощі у прогнозуванні імуногенності неоантигенів. Крім того, мікрооточення пухлини та імунна система також відіграють вирішальну роль у визначенні імуногенності неоантигенів. Тому необхідна експериментальна перевірка прогнозованих неоантигенів, щоб підтвердити їхню імуногенність і потенціал для використання в імунотерапії [22-24].

У класичному процесі ідентифікації неоантигенів пухлинні антигени визначаються за допомогою послідовності методів: 1) використання даних секвенування наступного покоління; 2) визначення пухлиноспецифічних

неоепітопів за допомогою прогнозів зв'язування in silico та тестів на імуногенність in vitro[25].

Але немає гарантії, що ці епітопи будуть представлені на пухлинних клітинах. Без присутності цільового епітопу трансформовані клітини не можуть бути ідентифіковані CD8+ Т-клітинами, що призводить до росту пухлини. З іншого боку, імунопептидоміка на основі мас-спектрометрії ідентифікує неоепітопи, які дійсно представлені і зв'язуються з HLA. Імунотерапія, націлена на ці епітопи, призводить до імунної відповіді CD8+ Т-клітин, що призводить до руйнування пухлини(Рис. 1.2) [26].

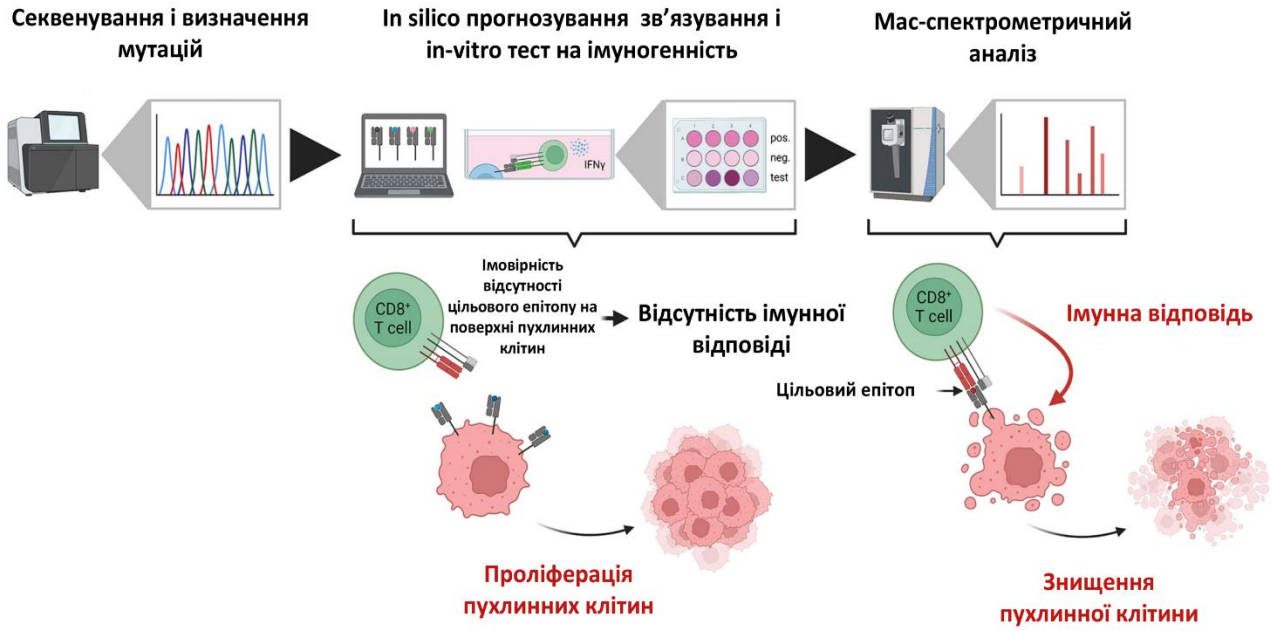

Рис. 1.2. Етапи класичного імунопептидомічного процесу ідентифікації цільового неоепітопу який одночасно присутній на поверні пухлинних клітин і здаден з достатньою афінністю зв'язуються з лейкоцитарним антигеном людини.

РОЗДІЛ 2

ОБ'ЄКТ, МЕТОДИ ТА МАТЕРІАЛИ ДОСЛІДЖЕННЯ

2.1. Об'єкт дослідження

Об'єктом цього дослідження є результати отримані лабораторіями Biomedical Informatics Lab університету ETH Zurich і OHSU Computational Bio університету Oregon Health & Science University (OHSU) з двох різних обчислювальних алгоритмів, заснованому на графах сплайсингу (GP) і сполученнях екзонів у сплайс-варіантах (JP) відповідно, які спрямовані на виявлення неоепітопів (9-мерів) трансльовваних із сполучення екзон-екзон.

2.2 Джерела даних

Дані про екзонні сполучення у сплайс-варіантах були обрані для когорти пухлинних зразків з проекту Атлас геному раку (TCGA) [27], і для когорти нормальних зразків з Проекту експресії генотипу (GTEx) відповідно [28]. Ці дані з експериментів RNA-seq, включають понад 11 000 зразків пухлин у 33 типах раку TCGA , 10 000 зразків, що охоплюють 29 нормальних типів тканин GTEx. Оприлюднена та уніфікована обробкас цих даних дозволили провести аналіз «нормального» та специфічного для раку сплайсингу. Використано файл анотації GENCODE v.32 і референсний геном GRCh38(hg38), завантажені відповідно з [29]. Протеом людини було завантажено з UniProt[30].

2.3 Відбір цільових зразків

Двома лабораторіями були створені набори сгенерованих даних неоепітопів, які включають інформацію по 19697 генам і 1102 зразкам пацієнтів з когорти раку молочної залози BRCA (TCGA) та когорти зразків нормальних тканин(молочної залози та всієї решти тканин включаючи імунопривілейовані такі як мозку, сім'янників, плаценти, очей) від здорових пацієнті у GTEx був створений набір даних нормальних когорт з 19697 генів і

9477 зразків пацієнтів.. На другому етапі обчислювань було загалом проведено 60 фільтраційних експериментів. У цьому досліджені випадковим чином обрано 5 цільових зразків і 5 фільтраційних параметрів.

2.4 Огляд інструментів у складі алгоритмів

Для вирівнювання зчитувань із еталонним геномом обома лабораторіями було використано безкоштовне програмне забезпечення STAR з відкритим вихідним кодом [31-32]. Для подальшої побудови графів сплайсингу для когорт нормальних і пухлинних зразків з даних RNA-seq та файлу анотацій GENCODE було використано інструмент SplAdder [33]. Код програми є у вільному доступі [34]. Даний інструмент перетворює анотації на відповідний граф сплайсингу, збагачуючи його доказами сплайсингу зі зразків секвенування, ідентифікація події сплайсингу з доповненого графу та використання даних секвенування для оцінки кількості подій сплайсингу (Рис. 2.1)[33]. Далі сгенерований у попередньому кроці граф сплайсингу приймає ImmunoPepper який є інструментом для генерації потенційних кандидатів неоепітопів з кодом у відкритому доступі [34]. Його суть полягає у генерації набору усіх можливих кандидатів послідовностей 9-мерів через проходження і трансляції усіх сполучень екзон-екзон у графі сплайсингу. Для подальшої фільтрації специфічних для раку неоепітопів можна використовувати відібрані 9-мери як вхідні дані для передбачення зв’язування MHC або використовувати бази даних мас-спектрометрії (MS) для подальшої валідації.

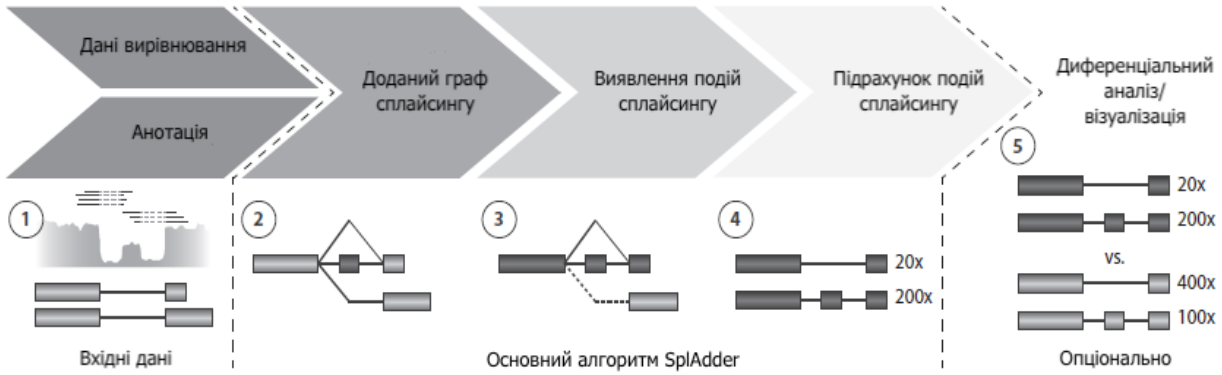

Рис. 2.1. Схематичне зображення алгоритму SplAdder який складається з 1-5 етапів.

2.5 Обчислювальні алгоритми GP і JP

Рис. 2.2. Огляд послідовності обчислювальних кроків для GP і JP алгоритмів.

У роботі порівнюються два концептуально ідентичних обчислювальні алгоритми для прогнозування специфічних для раку пептидів з 9 амінокислотних залишків.

Стадія генерації кандидатів неоепітопів (9-мерів):

1. Для підготовки вхідних даних порівнюють дані RNA-seq когорт пухлинних і нормальних зразків,використовуючи файл анотації генів у парі з референсним геномом людини;
2. Вирівнювання зчитувань RNA-seq із референсним геномом і кількісне визначення експресії за допомогою програмного забезпечення STAR;
3. У випадку GP передача даних вирівнювання в інструмент SplAdder, який трансформує анотації генів у відповідний граф сплайсингу(граф корти нормальних зразків з GTEx та ракових зразків з TCGA відповідно), збагачуючи його доказами сплайсингу із зразків

секвенування, вилучення всіх можливих пар сполучень екзон-екзон відповідно до графу, на відміну від цього, JP використовує на вході дані вирівнювання та виділяє всі бі-екзонні сполучення із подальшою інсерцією у транскрипт

4. Наступним кроком у випадку GP є трансляція пептидів шляхом застосування анотованих рамок зчитування як до анотованих, так і до нових транскриптів, або у випадку JP трансляція лише сполучень екзонів, чий передній сайт сплайсингу потрапляє в кодуючу область послідовності ДНК анотованого транскрипту, що кодує білок, у анотованій рамці зчитування → кмеризація пептидної послідовності навколо сполучення в 9-мери, що перекривають сполучення екзон-екзон. Нормалізація значень експресії всіх 9-мерних послідовностей виконується однаковим чином в обох алгоритмах, використовуючи метод “75-й квантиль нормалізації ”
5. Після цього GP і JP виконують звичайну фільтрацію кандидатів 9-мерів з нормальним протеомом людини UniProt. У випадку JP не були враховані параметри де амінокислоти лейцин (Leu) еквівалента ізолейцину (Ile).

Стадія фільтрації:

Кінцевою метою є віднімання специфічних для раку 9-мерів шляхом їх фільтрації проти 9-мерів присутніх у нормальній когорі. GP і JP виконують експерименти з однаковими наборами параметрів фільтрації,які будуть розглянуті у настпуному пункті.

Різниця у техніці виконання полягає у тому що, у GP 9-мери, що перекривають сполучення екзон-екзон, фільтруються проти всіх 9-мерів когорти, незалежно від наявності у них сполучення , а отже зосереджується на специфічних для раку пептидах, що виникають із сайтів сплайсингу, які можуть не бути новими або специфічними для раку. На відміну від цього, JP виконує фільтрацію сполучень екзон-екзон і ідентифікує пептиди, які транслюються безпосередньо з сайтів сплайсингу, характерних для раку.

Після цього наступними кроками майбутніх досліджень можуть стати перевірка протеоміки та прогноз афінності зв’язування пептиду з комплексом MHC.

2.6 Категорії фільтрів, застосовуваних на етапі фільтрації

Існує 4 незалежні фільтри:

- **Для корти пухлинних зразків** 1) ліміт експресії для цільового зразка > n (кількість reads) і 2) ліміт експресії для всіх зразків пухлинної когорти ≥ n (кількість reads) серед N – кількості зразків;
- **Для корти нормальних зразків** 3) ліміт експресії для всіх зразків пухлинної когорти ≥ n (кількість reads) серед N – кількості зразків;
- **Протеомна фільтрація**: вільтрування кандидатів неоепітопів проти у нормального протеому людини з UniProt.

Строгість правил фільтраціїї у **для корти пухлинних зразків** підвищує впевненість у точності прогнозованого сполучення сплайсингу або у трансльованому неоепітопі. Тоді як вимогливість у фільтрі **для корти нормальних зразків** підвищує впевненість у пухлинній специфічності утвореного неоепітопу.

2.7 Огляд інструментів аналізу і баз даних

2.7.1 Геномні бази даних

Для пошуку інформації про цікавлячі гени які є джерелом отриманих послідовностей 9-мерів та побудови біологічно релевантних гіпотез на основі отриманих результатів було використано Genome browser database для версії геному людини GRCh38(hg38) та такі бази як Ensembl database, The Human protein atlas.

2.7.2 Аналіз даних

В даній роботі була використана версія Python 3.10.2. обробки і аналізу даних, та побудови графіків.

2.8 Попередня обробка даних для аналізу

2.8.1 Порівняння наборів даних генерації неоепітопів зразків BRCA і GTEx та створення статистики

Було підраховано статистику унікальних послідовностей kmers (9-мерів) для усіх 19697 генів і 5 цільових зразків присутніх в наборах даних BRCA (ETH) і BRCA(OHSU) по окремості за допомогою методів set() і intersection(), а також кількість спільних 9-мерів, присутніх одночасно в обох наборах даних. Для GTEx було порівняно дані генерації рекурентних 9-мерів з певним лімітом експресії у всіх нормальних зразках.

Для перевірки з кількості спільних 9-мерів, який відсоток з них має походження зі спільних координат сполучення екзонів, а отже з однакових генів було сгенеровано статистику спільних координат між двома наборами даних.

І.Попередня обробка формату координат ETH:

Була виконана попередня обробка, щоб отримати лише координати, які знаходяться на межі екзонів, з повних координат пептидів з яких утворено 9-мери.

Функція (get_junction_coordinates) визначає порядок ланцюга “+” або “-” генетичної послідовності та обчислює координату сполучення екзон-екзон на основі кількості 2 або 3 екзони (Рис. 2.3) → для 2-екзонів із 4-значних координат 9-меру отримано 2-значні координати сполучення екзонів, а для 3-екзонів із 6-значних координат - 4-значні координати сполучення (Рис. 2.4)

```python
#ETH coordinates preprocessing:
def get_junction_coordinates(df, coordinates_col, sep=':'):
    df['strand'] = None
    df['junction_coordinate'] = None

    for idx, row in tqdm.tqdm(df.iterrows()):
        kmer_coordinates = [int(x) for x in row[coordinates_col].split(sep) if x !='None']

        if kmer_coordinates[1] < kmer_coordinates[2]: # order strand +

            df.loc[idx, 'strand'] = '+'
            if len(kmer_coordinates) == 4:  # 2 exons
                df.loc[idx, 'junction_coordinate'] = ':'.join([str(x) for x in kmer_coordinates[1:3]])
            elif len(kmer_coordinates) == 6:
                df.loc[idx, 'junction_coordinate'] = ':'.join([str(x) for x in kmer_coordinates[1:5]])
        else: # order strand +
            df.loc[idx, 'strand'] = '-'
            if len(kmer_coordinates) == 4:  # 2 exons
                df.loc[idx, 'junction_coordinate'] = ':'.join([str(x) for x in [kmer_coordinates[3],
                                                                                kmer_coordinates[0]]])
            elif len(kmer_coordinates) == 6:
                df.loc[idx, 'junction_coordinate'] = ':'.join([str(x) for x in [kmer_coordinates[3],
                                                                                kmer_coordinates[0],
                                                                                kmer_coordinates[2],
                                                                                kmer_coordinates[5]
                                                                               ]])
    return df
```

Рис. 2.3. Функція для отримання координат сполучення для ETH набору координат.

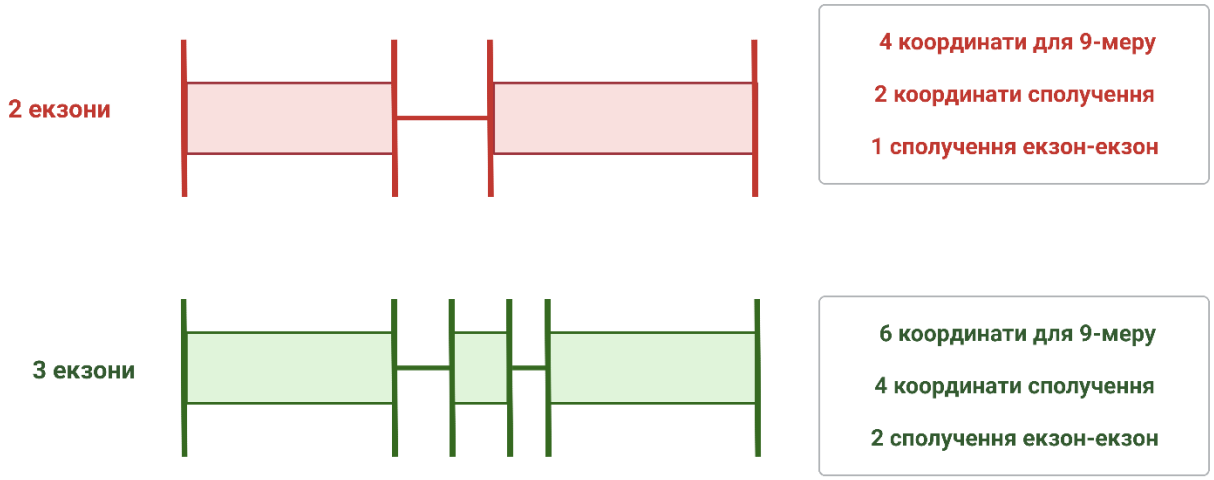

Рис. 2.4. Схематичне зображення транскрипту, складаючогося з 2 або 3 екзонів.

II.Попередня обробка формату координат OHSU:

Оскільки у наборі OHSU і ETH використані різні системи координат з UCSC Genome Browser, було прийняте рішення про переведення формату всіх координат OHSU 1-based* - координати, розміщені в веб-інтерфейсі UCSC Genome Browser у формат ETH координат 0-based* координати, що зберігаються в таблицях бази даних (Рис. 2.6.). Для цього було написано

функцію яка віднімає 1 від початку, а кінець координати залишає таким самим і потім вилучає координати сполучень екзонів для набору OHSU (Рис. 2.5).

```
#OHSU coordinates format changing:
def ohsu_to_eth_coord(df, col = 'jx', new_col = 'jx_shifted', sep = ';'):
    tmp_jx = df[col].str.split(sep,  expand = True)
    df[new_col] = tmp_jx[0] + sep + (tmp_jx[1].astype(int) - 1).astype(str) + sep + tmp_jx[2] + sep +
tmp_jx[3]
    return df
```

Рис. 2.5. Функція переводу формату координат сполучення екзонів для OHSU набору координат.

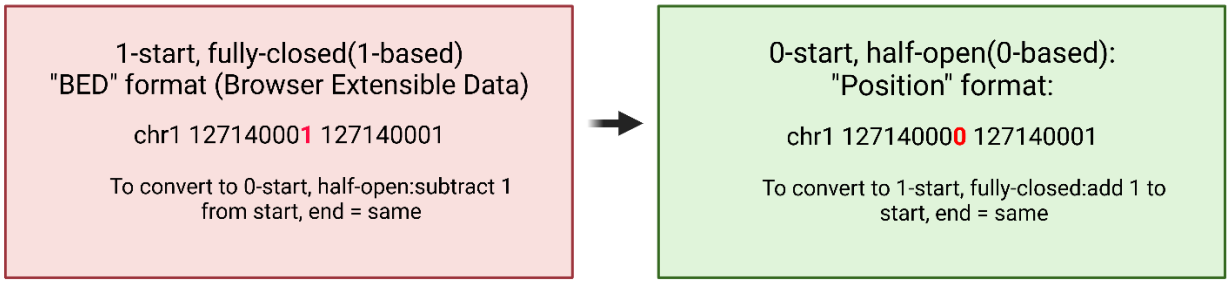

Рис. 2.6. Схематичне зображення переведення системи координат OHSU 1-based* у ETH 0-based*.

III. Нормалізація експресій послідовностей 9-мерів у зразках BRCA:

Метою подальшого аналізу було порівняння для 5 цільових зразків експресії спільних для наборів 9-мерів, які походять зі спільних генів.

Оскільки значення експресії у зразках неоепітопів набору даних OHSU було попередньо нормалізовано для їх подальшого порівняння із зразками ETH, було написано функцію для нормалізації експресії у зразках ETH (Рис. 2.4).

```
NORMALIZER_LIBSIZE = 400000
PATH_LIBSIZE =
'/cluster/work/grlab/projects/projects2020_OHSU/peptides_generation/CANCER_eth/commit_c4dd02c_conf2_Fra
me_cap0_runs/TCGA_Breast_1102/expression_counts.libsize.tsv'

def process_libsize(path_lib, custom_normalizer):
    '''
    Loads and returns the normalisation values per sample
    Normalisation formulae: (count / 75 quantile expression in sample) * A
    If no normalisation factor is provided: A = median across samples
    If normalisation factor is provided: A = normalisation factor
    :param path_lib: str path with library size file
    :param custom_normalizer: custum normalisation factor
    :return: dataframe with 75 quantile expression in sample / A values
    '''
    lib = pd.read_csv(path_lib, sep='\t')
    if custom_normalizer:
        lib['libsize_75percent'] = lib['libsize_75percent'] / custom_normalizer
    else:
        lib['libsize_75percent'] = lib['libsize_75percent'] / np.median(lib['libsize_75percent'])
    lib['sample'] = [sample.replace('-', '').replace('.', '').replace('_','') for sample in
lib['sample']]
    lib = lib.set_index('sample')
    return lib

def normalization(df, cols, libsize, metadata):
    df_expr = df.loc[:, cols]
    df_norm_vals = libsize.loc[cols, 'libsize_75percent']
    df = pd.concat([df.loc[:, metadata], df_expr / df_norm_vals], axis = 1)
    return df
```

Рис. 2.7. Функція для нормалізації експресії у зразках ETH.

В її основі **формула нормалізації експресії:**

$$(\text{count}/75\text{th quantile expression in sample}) * A \qquad (2.1)$$

де count — це кількість зчитувань, вирівняних до відповідного сполучення екзон-екзон, яке перекриває 9-мер і A – нормалізуючий фактор, середнє значення експресії для суми всіх 75-х значеннях експресій всіх зразків, що дорівнює 400000.

2.8.2 Порівняння даних фільтрації неоепітопів раку молочної залози (BRCA) та створення статистики

Було сгенеровано загальні таблиці статистики даних 5 зразків після застосування 60 різних фільтраційних параметрів, яка містить колонки з даними кількостей унікальних пухлиноспецифічних кандидатів 9-мерів в BRCA(ETH) і BRCA(OHSU) колонки size_eth і size_ohsu відповідно, size_ohsu\eth і size_eth\ohsu - колонки різниці кількості зразків 9-мерів між

двома наборами даних і size_intersection - колонка спільних 9-мерів для eth і ohsu зразків (Рис. 2.8 – 2.12).

Filtered data of sample TCGAC8A12P01A11RA11507all					
Filter #	size_ohsu	size_eth	size_intersection	size_ohsu\eth	size_eth\ohsu
1	7287	895	125	7162	770
2	2336	812	330	2006	482
3	553	733	172	381	561
4	7388	895	126	7262	769
5	2754	883	395	2359	488

Рис. 2.8. Таблиця статистики кількостей унікальних і спільних 9-мерів для зразка TCGAC8A12P01A11RA11507.

Filtered data of sample TCGAAOA0JM01A21RA05607all					
Filter #	size_ohsu	size_eth	size_intersection	size_ohsu\eth	size_eth\ohsu
1	3643	1071	312	3331	759
2	907	577	163	744	414
3	390	458	98	292	360
4	3690	1071	312	3378	759
5	1155	607	238	917	369

Рис. 2.9. Таблиця статистики кількостей унікальних і спільних 9-мерів для зразка TCGAAOA0JM01A21RA05607.

Filtered data of sample TCGABHA18V01A11RA12D07all					
Filter #	size_ohsu	size_eth	size_intersection	size_ohsu\eth	size_eth\ohsu
1	4860	1136	153	4707	982
2	2882	1038	350	2532	688
3	584	954	151	433	803
4	4932	1135	157	4775	978
5	3421	1194	483	2938	711

Рис. 2.10. Таблиця статистики кількостей унікальних і спільних 9-мерів для зразка TCGABHA18V01A11RA12D07.

Filtered data of sample TCGAA2A0D201A21RA03407all					
Filter #	size_ohsu	size_eth	size_intersection	size_ohsu\eth	size_eth\ohsu
1	4545	872	82	4463	790
2	2404	838	296	2108	542
3	547	764	141	406	623
4	4608	878	108	4500	770
5	2886	934	410	2476	524

Рис. 2.11. Таблиця статистики кількостей унікальних і спільних 9-мерів для зразка TCGAA2A0D201A21RA03407.

Filtered data of sample TCGAA2A0SX01A12RA08407all					
Filter #	size_ohsu	size_eth	size_intersection	size_ohsu\eth	size_eth\ohsu
1	4350	1018	205	4145	813
2	2343	832	288	2055	544
3	483	737	131	352	606
4	4441	1028	214	4227	814
5	2792	912	398	2394	514

Рис. 2.12. Таблиця статистики кількостей унікальних і спільних 9-мерів для зразка TCGAA2A0SX01A12RA08407.

РОЗДІЛ 3

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ ТА ЇХ ОБГОВОРЕННЯ

3.1 Результати порівняння статистики координат сполучень екзонів

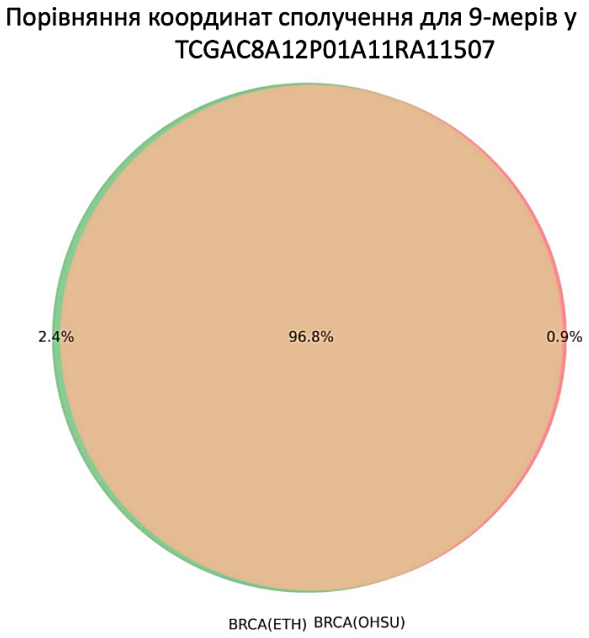

Рис. 3.1. Діаграма Вена з даними унікальних і спільних координат сполучень екзонів для кандидатів 9-мерів у зразку TCGAC8A12P01A11RA11507.

Рис. 3.2. Діаграма Вена з даними унікальних і спільних координат сполучень екзонів для кандидатів 9-мерів у зразку TCGAAOA0JM01A21RA05607.

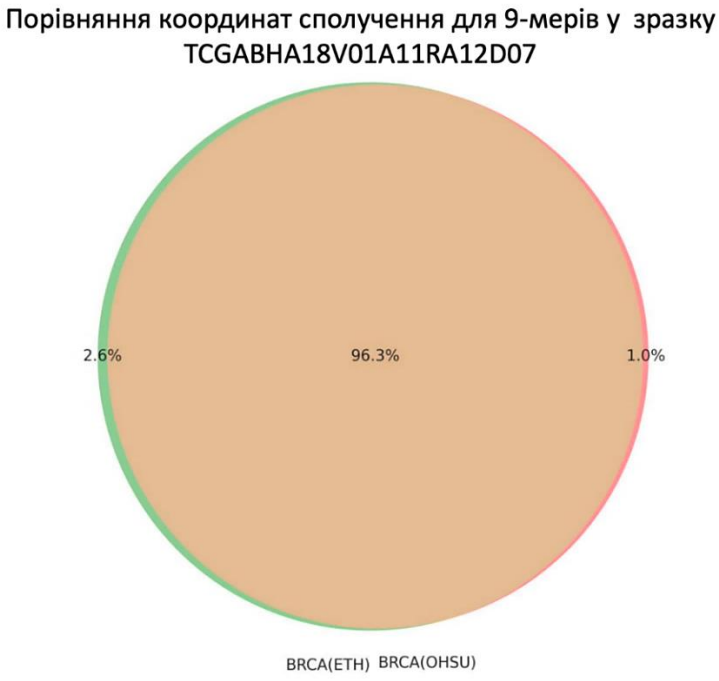

Рис. 3.3. Діаграма Вена з даними унікальних і спільних координат сполучень екзонів для кандидатів 9-мерів у зразку TCGABHA18V01A11RA12D07.

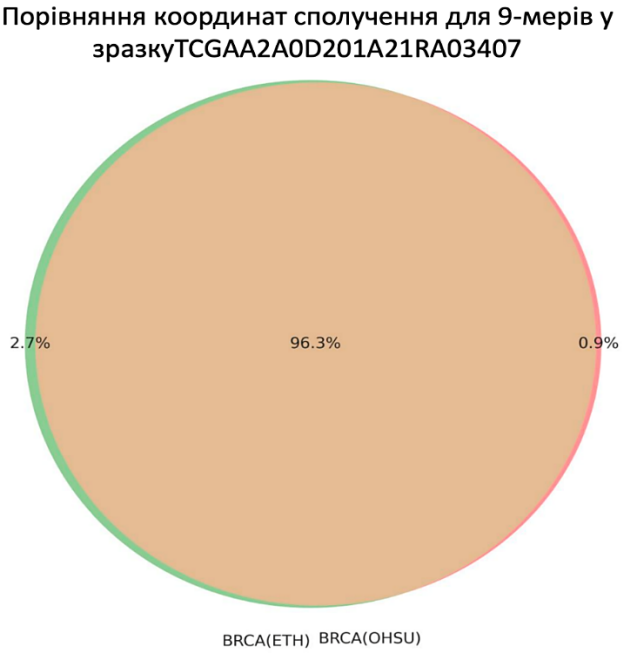

Рис. 3.4. Діаграма Вена з даними унікальних і спільних координат сполучень екзонів для кандидатів 9-мерів у зразку TCGAA2A0D201A21RA03407.

Рис. 3.5. Діаграма Вена з даними унікальних і спільних координат сполучень екзонів для кандидатів 9-мерів у зразку TCGAA2A0SX01A12RA08407.

На діагармах Вена (Рис. 3.1. – 3.5) великий відсоток співпадіння значень спільні дані пар однакових 9-мерів і координат у наборах даних BRCA(ETH) і BRCA(OHSU) з двох алгоритмів GP і JP відповідно доводить значну, але не абсолютну відтворюваність їх результатів.

Однією з причин, чому це не 100% перекриття, може бути різниця в техніці генерації координат. У випадку алгоритму GP після трансляції пар біекзонів і їх розрізання на 9-мери, у випадку, коли довжина амінокислоти другого екзона коротша за довжину 9-мерів, пептид біекзону було розширено за рахунок додавання справа третього екзону. В результаті ми маємо в наборі даних ETH (BRCA) ті однакові 9-мери які походять з одного і того самого гену але їх координати сполучення екзонів складаються з 4 значень для 3 екзонів.

3.2 Результати порівняльного аналізу велечин експресій спільних 9-мерів в 5 цільових зразках

Рис. 3.6. Стовпчикова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGAA2A0SX01A12RA08407 і гену ENSG00000162576.16.

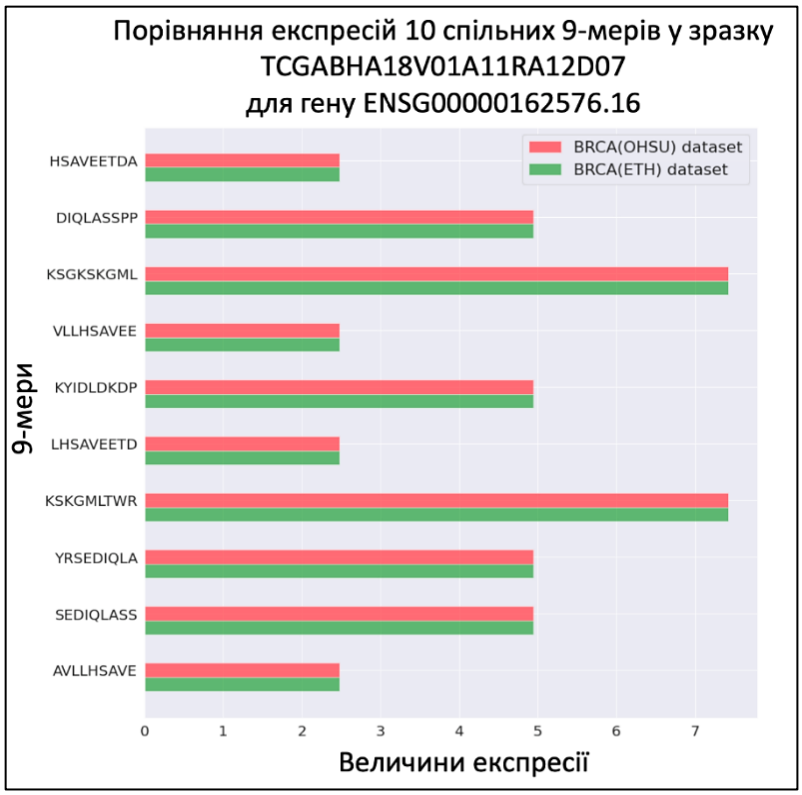

Рис. 3.7. Стовпчикова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGABHA18V01A11RA12D07 і гену ENSG00000162576.16.

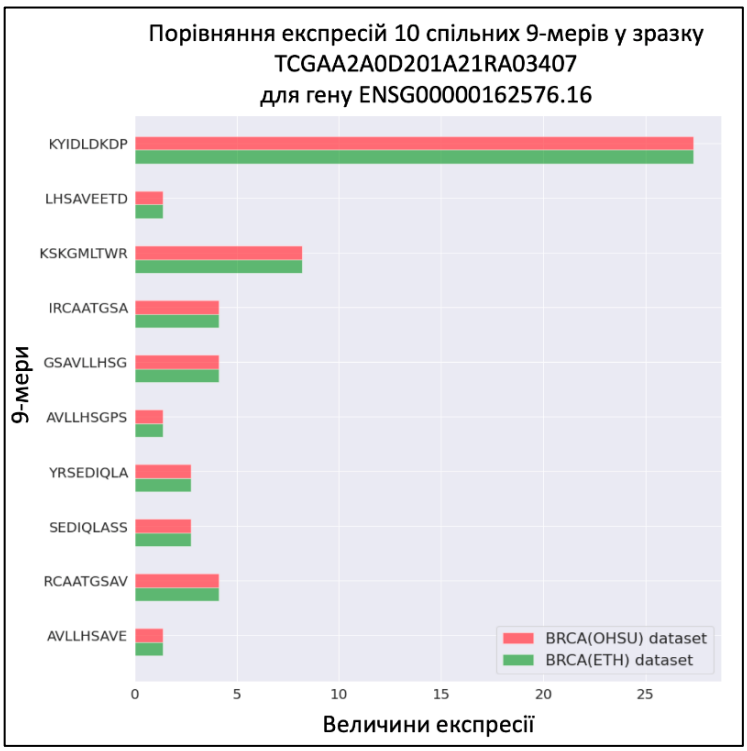

Рис. 3.8. Стовпчикова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGAA2A0D201A21RA03407 і гену ENSG00000162576.16.

Рис. 3.9. Стовпчикова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGAAOA0JM01A21RA05607 і гену ENSG00000162576.16.

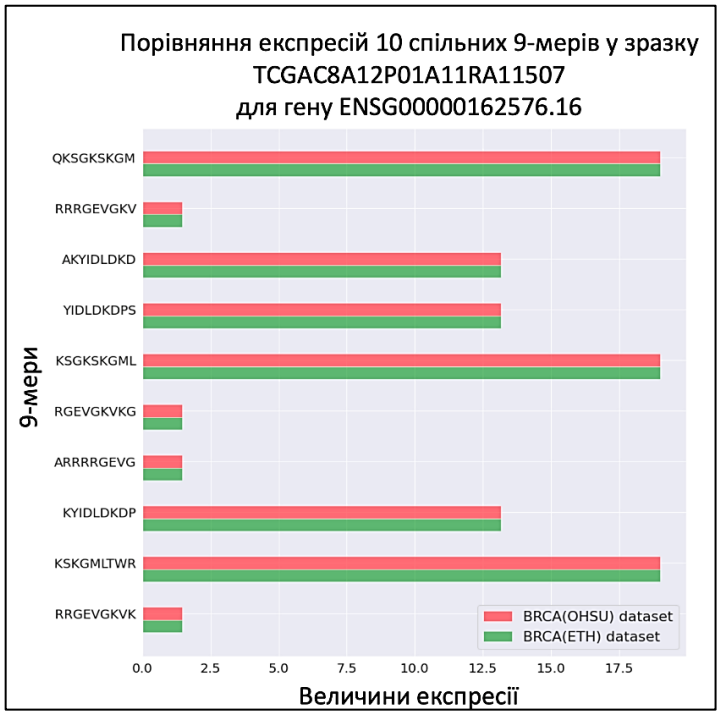

Рис. 3.10. Стовпчикова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGAC8A12P01A11RA11507 і гену ENSG00000162576.16.

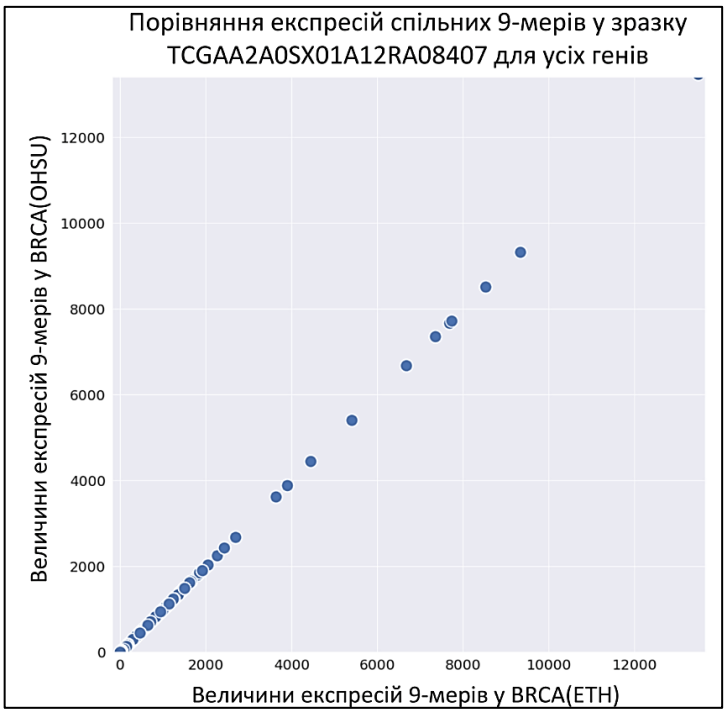

Рис. 3.11. Точкова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGAA2A0SX01A12RA08407 і всіх генів.

Рис. 3.12. Точкова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGAC8A12P01A11RA11507 і всіх генів.

Рис. 3.13. Точкова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGABHA18V01A11RA12D07 і всіх генів.

Рис. 3.14. Точкова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGAA2A0D201A21RA03407 і всіх генів.

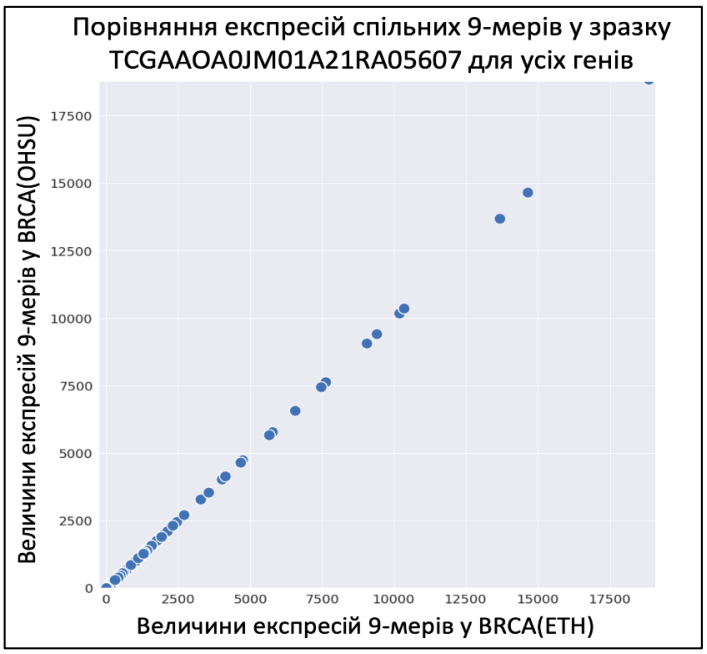

Рис. 3.15. Точкова діаграма порівняння значень експресії для однакових кандидатів 9-мерів між даними BRCA(ETH) і BRCA(OHSU) зразку TCGAAOA0JM01A21RA05607 і всіх генів.

Узгодження значень експресії спільних 9-мерів між наборами даних ETH і OHSU (BRCA) на стовпчикових діаграмах (Рис.3.6 -3.10) та точкових діагарамх (Рис. 3.11 – 3.15) підтверджують надійність кількісної оцінки експресії. Це означає, що обчислення експресії (підрахунок прочитаних послідовностей, пов'язаних з генами) при вирівнюванні в програмі STAR було виконано правильно та відтворювано в обох методах (кожна координата сполучення екзон-екзон пов’язана з кількістю вирівняних первинних зчитувань, перекриваючих область сполучення екзонів).

3.3 Порівняння даних генерації неоепітопів зі здорових зразків GTEx

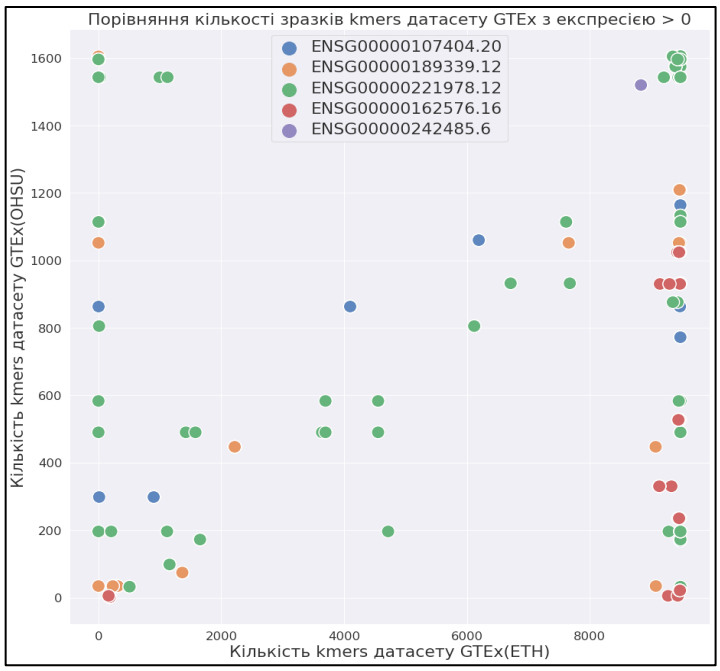

Рис. 3.16. Точкова діаграма порівняння кількостей однакових кандидатів 9-мерів з експресією >0 у всіх зразках для 5 генів між GTEx(ETH) і GTEx(OHSU).

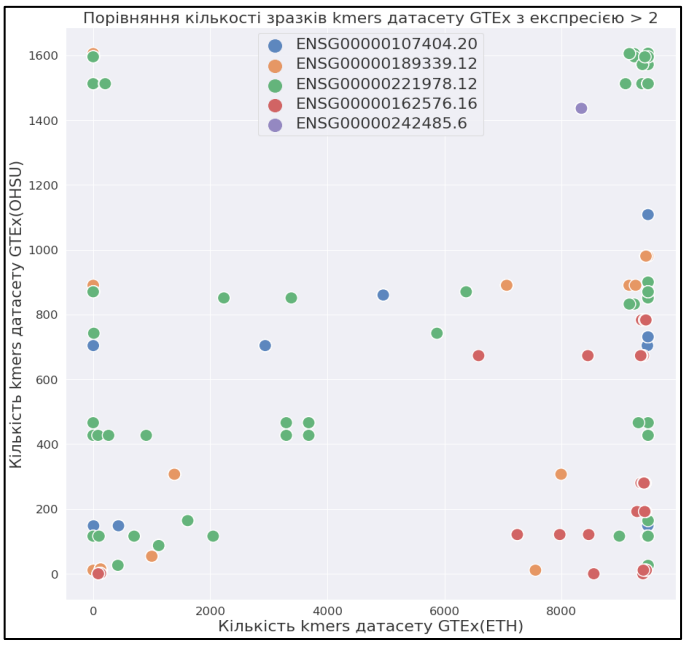

Рис. 3.17. Точкова діаграма порівняння кількостей однакових кандидатів 9-мерів з експресією >2 у всіх зразках для 5 генів між GTEx(ETH) і GTEx(OHSU).

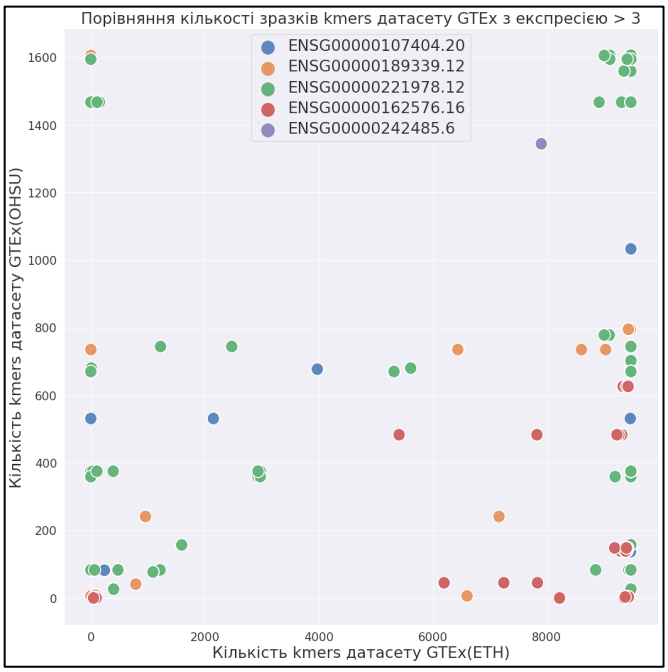

Рис. 3.18. Точкова діаграма порівняння кількостей однакових кандидатів 9-мерів з експресією >3 у всіх зразках для 5 генів між GTEx(ETH) і GTEx(OHSU).

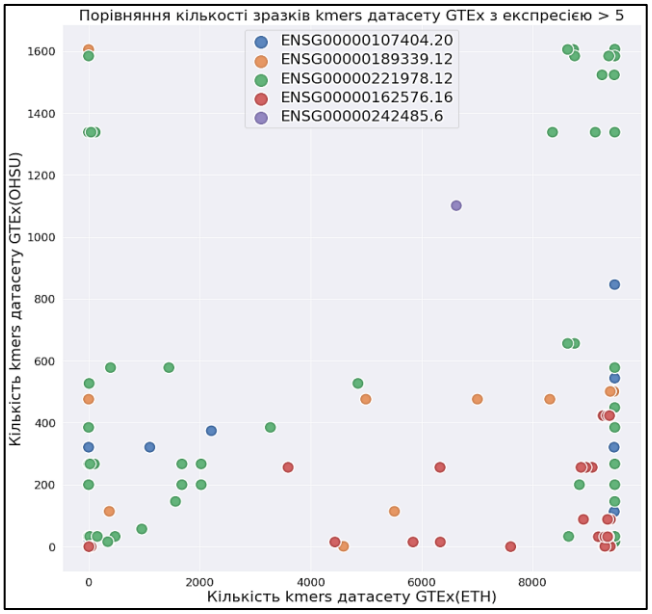

Рис. 3.19. Точкова діаграма порівняння кількостей однакових кандидатів 9-мерів з експресією >5 у всіх зразках для 5 генів між GTEx(ETH) і GTEx(OHSU).

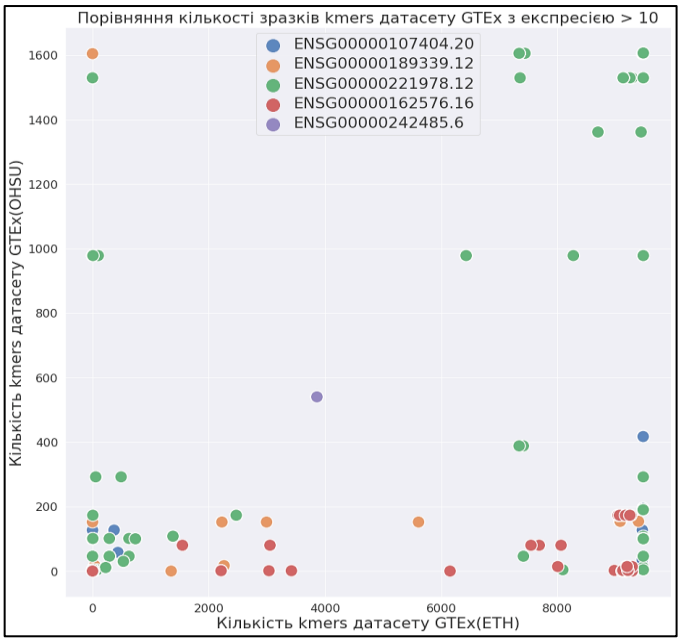

Рис. 3.20. Точкова діаграма порівняння кількостей однакових кандидатів 9-мерів з експресією >10 у всіх зразках для 5 генів між GTEx(ETH) і GTEx(OHSU).

Велика різниця в кількості 9-мерів між наборами даних ETH(GTEx) і OHSU(GTEx) на точкових діаграмах (Рис. 3.16 - 3.20) ставить під сумнів повноту набору даних OHSU. Можливим поясненням є меньший набір даних нормальних когортних зразків(7000) OHSU, на противагу у ETH 9477 зразків і тому було помилково розраховано рекурентність 9-мерів з різними лімітами експресіїї >0, >2, >3, >5, >10 у всіх зразках нормальної когорти. Вирішенням цієї помилки є повторна генерація кандидатів алгоритмом JP.

3.4 Результати порівняння пухлиноспецифічних неоепітопів раку молочної залози (BRCA) після застосування 5 різних фільтрів

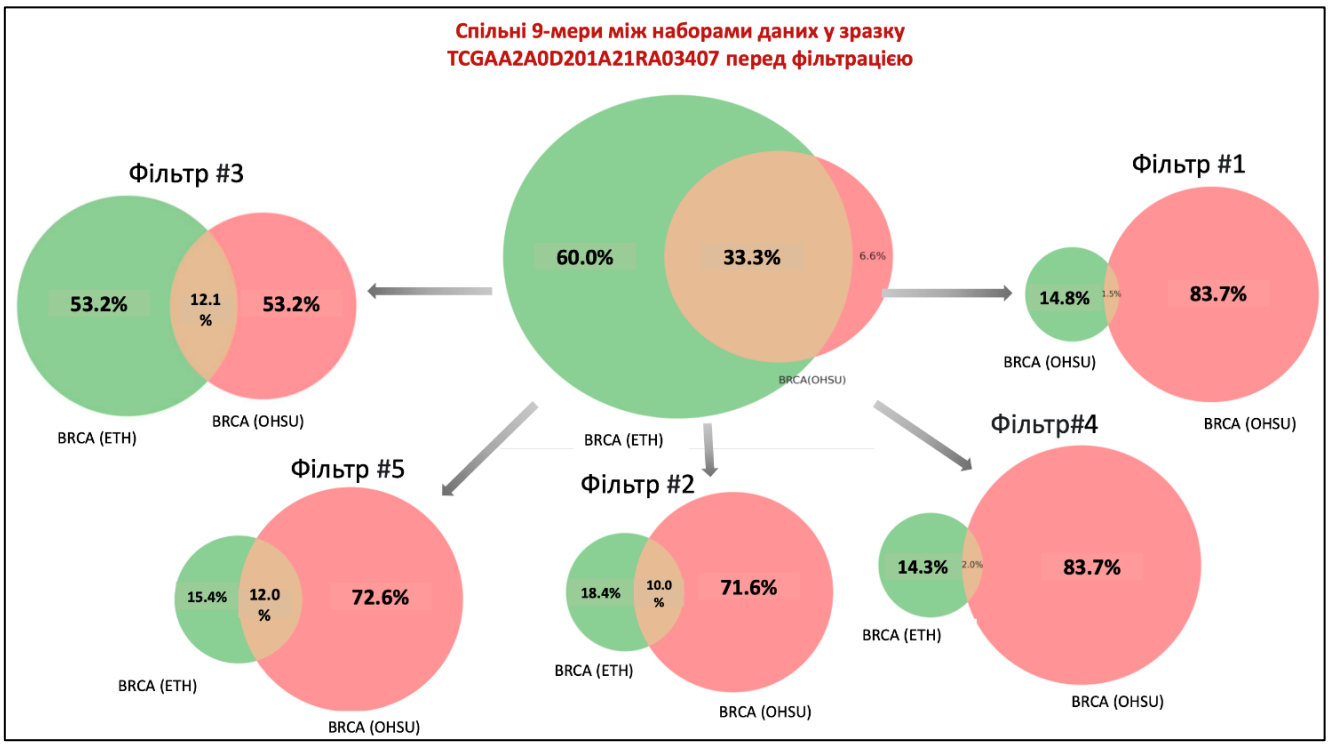

Рис. 3.21. Порівняння даних спільних посідовностей 9-мерів між двома наборами даних BRCA(ETH) і BRCA(OHSU) для зразка до і після застосування 5 різних фільтрів.

Діаграми Вена на Рис. 3.21. показують різні ефекти фільтрації для одного зразка. Аналогічні графіки було отримано для решти цільових зразків. Фільтри №3 і №5 демонструють найбільші області перетину даних потенційно специфічних для раку 9-мерів.

У фільтрі № 3 це можна пояснити застосуванням суворих правил щодо 9-мерів у нормальній когорті (9-мери повинні мати експресію ≥ 3 у 10 нормальних зразках), у той же час параметри фільтра для когорти раку не строгі (9-мер у зразку інтересу повинен мати >0 зчитувань і ≥ 2 зчитувань в 1 когортному зразку раку). Це означає, що тоді видаляється менша частина 9-мерів присутніх у нормальній когорті із 9-мерів зразків раку молочної залози. Схожий ефект спостережено для **фільтру №5**: параметри фільтру для нормальних зразків є більш строгими порівняно з параметрами до зразків

пухлини, де експресія цільового та цілої когорти має бути > 0 у будь-якій кількості зразків, для нормальних зразків базова повторюваність становить ≥3 для будь-якої кількості зразків.

У **фільтрах №2, №4 і №1** протилежна ситуація: нормальний когортний фільтр не є суворим (вимогою щодо включення у вибірку 9-мерів є експресія >0 в 1 зразку), водночас у №1 фільтр обмеження експресії для ракової когорти не застосовується, тому кількість кандидатів неоепітопів залишається незмінною, але при цьому видаляється більший розмір нормальних 9-мерів із ракових 9-мерів.

До неочікуваних результатів можна віднести більшу частку відфільтрованих 9-мерів, ідентифікованих алгоритмом JP, що ставить під сумнів точність протеомної фільтрації проти нормального протеому людини з UniProt на етапі генераціїї кандидатів неоепітопів.

Одним із пояснень може бути відсутність еквівалентів лейцину та ізолейцину, заміни символів амінокислоти ізолейцину на лейцини не була виконана алгоритмом JP у фільтрації з UniProt. У майбутній роботі це може стати перешкодою у валідації за допомогою MS, оскільки вони є нерозрізненими через ідентичні молекулярні маси, тому заплановано переробити частину фільтрації і повторити порівняльний аналіз даних GP і JP.

Для того щоб робити будь-які припущення щодо біологічної значимості відібраних кандидатів потенційних пухлиноспецифічних неоантигенів спільних для обох алгоритмів у майбутньому знадобляться подальші прогнози зв’язування з комплексом MHC in silico та тести на імуногенність in vitro. Основною вимогою для ідентифікації дійсних пухлинних мішеней є їх представленність на клітинах. Тому потрібна валідація за допомогою імунопептидоміки, а саме першочергова перевірка методом мас-спектрометрії. Використання таких неоепітопів в якості мішеней

в імунотерапії має призвести до імунної атаки CD8+ Т-клітинами та руйнування пухлини.

3.5 Біологічна релевантність результатів

Результатом аналізу з використанням геномних баз даних є таблиці для 5 зразків (Додаток 1) з описом кодуючих білок генів походження для деяких відфільтрованих специфічних для раку 9-мерів. Загалом більшість з перевірених 9-мерів пов’язані безпосередньо з раком молочної залози, або мають низьку ракову специфічність, тобто зустрічаються в інших типах. У решті випадків гени походження відносяться до конститутивних генів домашнього господарства, необхідних для підтримки основних клітинних функцій.

Одним із обмежень цієї частини роботи був малий розмір проаналізованого набору даних (5 цільових зразків із загальної 1102 к-ті зразків BRCA), тобто відсутність статистичної значущості для побудови гіпотез.

Оскільки кінцевою метою аналізу було знайти будь-яку біологічну релевантну інформацію, а саме дати відповіді на запитання: 1) чи для всіх кандидатів неоепітопів ген походження є пухлиноспецифічним; 2) чи існує кореляція між кількість подій сплайсингу (варіантів сплайсингу) і відношення гену до виникнення раку.

Вирішенням даного обмеження у майбутньому може бути комплексний аналіз повних наборів даних із застосуванням інструментів аналізу збагачення генної онтології на Python, GOATOOLS для систематизації і подальшої біологічної інтерпретації списків генів із цікавими спільними властивостями.

ВИСНОВКИ

1) Деякі з результатів аналізу потенційно можуть бути використані для підвищення точності обох алгоритмів.

2) Подальшою метою майбутніх досліджень є аналіз повного набору даних усіх зразків 1102 BRCA та 60 параметрів фільтрації для створення повної статистичної та біологічної інтерпретації.

3)Кінцевою метою подальших досліджень є перевірка імуногенності потенційних пухлиноспецифічних кандидатів шляхом валідації експресії неоепітопів за допомогою мас-спектрометрії і внутрішньоклітинних даних протеоміки зразків з CPTAC і перевірки афінності зв'язування з комплексом MHC за допомогою інструменту NetMHC.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Blass, E., Ott, P.A. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nat Rev Clin Oncol* **18**, 215–229 (2021). https://doi.org/10.1038/s41571-020-00460-2.
2. Coventry, B. J. & Henneberg, M. The immune system and responses to cancer: Coordinated evolution. *F1000Res.* **4**, 552 (2021).
3. Yarchoan, M., Johnson, B., Lutz, E. et al. Targeting neoantigens to augment antitumour immunity. Nat Rev Cancer 17, 209–222 (2017).
4. Aleksic, M. et al. Different affinity windows for virus and cancer-specific T-cell receptors: implications for therapeutic strategies. *Eur. J. Immunol.* 42, 3174–3179 (2012).

5. Nurieva, R., Wang, J. & Sahoo, A. T-cell tolerance in cancer. Immunotherapy **5**, 513–531 (2013).
6. Marei HE, Hasan A, Pozzoli G, Cenciarelli C. Cancer immunotherapy with immune checkpoint inhibitors (ICIs): potential, mechanisms of resistance, and strategies for reinvigorating T cell responsiveness when resistance is acquired. Cancer Cell Int. 2023 Apr 10;23(1):64. doi: 10.1186/s12935-023-02902-0. PMID: 37038154; PMCID: PMC10088229.
7. Zhang JY, Looi KS, Tan EM. Identification of tumor-associated antigens as diagnostic and predictive biomarkers in cancer. Methods Mol Biol. 2009;520:1-10. doi: 10.1007/978-1-60327-811-9_1. PMID: 19381943; PMCID: PMC2839120.
8. Jiang, T., Shi, T., Zhang, H. *et al.* Tumor neoantigens: from basic research to clinical applications. *J Hematol Oncol* **12**, 93 (2019).
9. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015 Apr 3;348(6230):69-74. doi: 10.1126/science.aaa4971. PMID: 25838375.
10. Hacohen N, Fritsch EF, Carter TA, Lander ES, Wu CJ. Getting personal with neoantigen-based therapeutic cancer vaccines. Cancer Immunol Res. 2013 Jul;1(1):11-5. doi: 10.1158/2326-6066.CIR-13-0022. Epub 2013 Apr 7. PMID: 24777245; PMCID: PMC4033902.
11. Sim MJW, Sun PD. T Cell Recognition of Tumor Neoantigens and Insights Into T Cell Immunotherapy. Front Immunol. 2022 Feb 10;13:833017. doi: 10.3389/fimmu.2022.833017. PMID: 35222422; PMCID: PMC8867076.
12. Adams, M.D.; Kerlavage, A.R.; Fleischmann, R.D.; Fuldner, R.A.; Bult, C.J.; Lee, N.H.; Kirkness, E.F.; Weinstock, K.G.; Gocayne, J.D.; White, O.; et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 1995, 377, 3–174.
13. Wang BD, Lee NH. Aberrant RNA Splicing in Cancer and Drug Resistance. Cancers (Basel). 2018 Nov 20;10(11):458. doi: 10.3390/cancers10110458. PMID: 30463359; PMCID: PMC6266310.

14. Fackenthal JD, Godley LA. Aberrant RNA splicing and its functional consequences in cancer cells. Dis Model Mech. 2008 Jul-Aug;1(1):37-42. doi: 10.1242/dmm.000331. PMID: 19048051; PMCID: PMC2561970.
15. Graveley BR. Sorting out the complexity of SR protein functions. RNA. 2000 Sep;6(9):1197-211. doi: 10.1017/s1355838200000960. PMID: 10999598; PMCID: PMC1369994.
16. Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, Potter NE, Heuser M, Thol F, Bolli N, Gundem G, Van Loo P, Martincorena I, Ganly P, Mudie L, McLaren S, O'Meara S, Raine K, Jones DR, Teague JW, Butler AP, Greaves MF, Ganser A, Döhner K, Schlenk RF, Döhner H, Campbell PJ. Genomic Classification and Prognosis in Acute Myeloid Leukemia. N Engl J Med. 2016 Jun 9;374(23):2209-2221. doi: 10.1056/NEJMoa1516192. PMID: 27276561; PMCID: PMC4979995.
17. Gimeno-Valiente F, López-Rodas G, Castillo J, Franco L. Alternative Splicing, Epigenetic Modifications and Cancer: A Dangerous Triangle, or a Hopeful One? Cancers (Basel). 2022 Jan 22;14(3):560. doi: 10.3390/cancers14030560. PMID: 35158828; PMCID: PMC8833605.
18. Zhang, Y., Qian, J., Gu, C. *et al.* Alternative splicing and cancer: a systematic review. *Sig Transduct Target Ther* **6**, 78 (2021).
19. Kishor A, Fritz SE, Hogg JR. Nonsense-mediated mRNA decay: The challenge of telling right from wrong in a complex transcriptome. Wiley Interdiscip Rev RNA. 2019 Nov;10(6):e1548. doi: 10.1002/wrna.1548. Epub 2019 May 26. PMID: 31131562; PMCID: PMC6788943.
20. Nellore, A. et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol. 17, 266(2016).
21. Kahles, A. et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell 34, 211–224.e6 (2018).
22. Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, Audehm S, Straub M, Weber J, Slotta-Huspenina J, Specht K, Martignoni ME, Werner A, Hein

R, H Busch D, Peschel C, Rad R, Cox J, Mann M, Krackhardt AM. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat Commun. 2016 Nov 21;7:13404. doi: 10.1038/ncomms13404. PMID: 27869121; PMCID: PMC5121339.

23. Yarchoan M, Johnson BA 3rd, Lutz ER, Laheru DA, Jaffee EM. Targeting neoantigens to augment antitumour immunity. Nat Rev Cancer. 2017 Apr;17(4):209-222. doi: 10.1038/nrc.2016.154. Epub 2017 Feb 24. Erratum in: Nat Rev Cancer. 2017 Aug 24;17 (9):569. PMID: 28233802; PMCID: PMC5575801.
24. van Rooij, N. et al. Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab- responsive melanoma. *J. Clin. Oncol.* **31**, e439–e442 (2013).
25. Hadrup SR, Bakker AH, Shu CJ, Andersen RS, van Veluw J, Hombrink P, Castermans E, Thor Straten P, Blank C, Haanen JB, Heemskerk MH, Schumacher TN. Parallel detection of antigen-specific T-cell responses by multidimensional encoding of MHC multimers. Nat Methods. 2009 Jul;6(7):520-6. doi: 10.1038/nmeth.1345. Epub 2009 Jun 21. PMID: 19543285.
26. Carreno BM, Magrini V, Becker-Hapak M, Kaabinejadian S, Hundal J, Petti AA, et al. Cancer Immunotherapy. A Dendritic Cell Vaccine Increases the Breadth and Diversity of Melanoma Neoantigen-Specific T Cells. Science (New York NY) (2015) 348(6236):803–8.
27. Cancer Genome Atlas Research Network; Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113-20. doi: 10.1038/ng.2764. PMID: 24071849; PMCID: PMC3919969.
28. Lonsdale, J., Thomas, J., Salvatore, M. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).
29. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, Berry A, Bignell A, Boix C, Carbonell Sala S, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C,

Gonzalez JM, Grego T, Hardy M, Hourlier T, Howe KL, Hunt T, Izuogu OG, Johnson R, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Riera FC, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Wolf MY, Xu J, Yang YT, Yates A, Zerbino D, Zhang Y, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Tress ML, Flicek P.Nucleic Acids Res 2021 : 49 ; d1 ; D916-D923.PUBMED: 33270111; PMC: PMC7778937; DOI: 10.1093/nar/gkaa1087

30. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res.47, D506–D515 (2019).
31. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21(2013).
32. Nellore, A. et al. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. Bioinformatics 33, 4033–4040 (2017).
33. Kahles, A., Ong, C. S., Zhong, Y. & Rätsch, G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics 32, 1840–1847(2016).
34. https://github.com/ratschlab/spladder
35. https://github.com/ratschlab/immunopepper

ДОДАТКИ

ДОДАТОК 1(Опис неоепітопів із зразка TCGAC8A12P01A11RA11507)

Послідовність неоепітопу (kmer, k=9)	Білок-кодуючий ген походження	Варіанти сплайсингу	Експресований у	Зв'язок з раком
GALVYAAKP	ENSG00000135413.9 (LACRT)	4 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** **рак молочної залози** - функція невідома (в основному) **Специфічність клітинної лінії**: Не виявлено	**Специфічний для раку:** **раку молочної залози** **BRCA** **Прогнозне резюме:** Генний продукт не є прогностичним
LRSSRDKTY	ENSG00000108588.14 (CCDC47)	6 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** **рак молочної залози** - функція невідома (в основному) **Специфічність клітинної лінії:** Низька специфічність раку	**Специфічний для раку:** **Низька специфічність раку** **Прогнозне резюме:** Прогностичний маркер колоректального раку (сприятливий)
VHLVTHDAR	ENSG00000198034.11 (RPS4X)	5 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічний - трансляція (переважно) **Специфічність одного типу клітин:** **Міоепітеліальні клітини молочної залози, залозисті клітини молочної залози** **Специфічність клітинної лінії:** Низька специфічність раку **Тканинна специфічність:**	**Специфічний для раку:** **Низька специфічність раку** **Прогнозне резюме:** Прогностичний маркер раку нирки (несприятливий) і раку щитовидної залози (сприятливий)

			Низька тканинна специфічність	
VHQHTSRSW	ENSG00000083857.14 (FAT1)	11 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Клітини сполучної тканини - організація ECM (переважно) **Специфічність клітинної лінії:** Низька специфічність раку **Класифікація типів клітин тканини:** **Збагачений тип клітин (молочна залоза - міоепітеліальні клітини молочної залози)**	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер при раку легені (несприятливий)

ДОДАТОК 2(Опис неоепітопів із зразка TCGAAOA0JM01A21RA05607)

Послідовність неоепітопу (kmer, k=9)	Білок-кодуючий ген походження	Варіанти сплайсингу	Експресований у	Зв’язок з раком
ASRRSPSGS	ENSG00000167257.1 (RNF214)	8 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічний - Основні клітинні процеси (переважно) **Специфічність клітинної лінії**: Низька специфічність раку	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер при раку печінки (несприятливий)
TFGKVQMSL	ENSG00000196460.14 (RFX8)	6 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Мієлоїдні клітини - вроджена імунна відповідь (переважно) **Специфічність**	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Генний продукт не є прогностичним

			клітинної лінії: Посилення раку (лейкемія, саркома) **Тканинна специфічність**: Тканинний посилений (кістковий мозок) **Кластер експресії тканин:** Кістковий мозок - сплайсинг мРНК і клітинний цикл (в основному)	
HTKQLASRR	ENSG00000167257.11 (RNF214)	8 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічний - трансляція (переважно) **Специфічність одного типу клітин:** Покращений тип клітин (сперматоцити) **Специфічність клітинної лінії:** Низька специфічність раку **Тканинна специфічність:** Низька тканинна специфічність	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер при раку печінки (несприятливий)

ДОДАТОК 3(Опис неоепітопів із зразка TCGABHA18V01A11RA12D07)

Послідовність неоепітопу (kmer, k=9)	Білок-кодуючий ген походження	Варіанти сплайсингу	Експресований у	Зв’язок з раком

ISVVNHQDH	ENSG00000152558.15 (TMEM123)	8 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Кератиноцити - функція епітеліальних клітин (в основному) **Специфічність клітинної лінії**: Низька специфічність раку	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер при раку підшлункової залози (несприятливий) і меланомі (несприятливий)
LHGPGLPRT	ENSG00000205476.9 (CCDC85C)	7 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічна – невідома функція (переважно) **Специфічність клітинної лінії:** Низька специфічність раку	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер раку нирки (сприятливий) та раку ендометрію (несприятливий)
ELQSQHLFF	ENSG00000078114.18 (NEBL)	24 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічний - трансляція (переважно) **Специфічність одного типу клітин:** Група збагачена (кардіоміоцити, астроцити) **Специфічність клітинної лінії:** Cancer enhanced (Rhabdoid) **Тканинна специфічність:** Збагачена тканина (серцевий м'яз)	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер раку нирки (сприятливий) та уротеліального раку (несприятливий)

RTEPSPNRV	ENSG00000143797.12 (MBOAT2)	10 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічна – невідома функція (переважно) **Специфічність клітинної лінії:** Низька специфічність раку	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер при раку ендометрію (несприятливий) та раку уротелію (несприятливий)

ДОДАТОК 4 (Опис неоепітопів із зразка TCGAA2A0D201A21RA03407)

Послідовність неоепітопу (kmer, k=9)	Білок-кодуючий ген походження	Варіанти сплайсингу	Експресований у	Зв’язок з раком
TGLCQIFSE	ENSG00000237441.10 (RGL2)	11 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** **Неспецифічний - Основні клітинні процеси (переважно)** **Специфічність клітинної лінії**: Низька специфічність раку **Одноклітинний тип кластер експресії:** груди - лактація (переважно)	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер при раку підшлункової залози (сприятливий)
QIMKGGKRS	ENSG00000131459.13 (GFPT2)	8 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічна – невідома функція (переважно) **Специфічність клітинної лінії:** Низька специфічність раку	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер раку нирки (несприятливий), раку легенів (несприятливий), раку щитовидної залози

				(несприятливий) та раку шийки матки (несприятливий)
LKLSAECQK	ENSG00000141627.13 (DYM)	20 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічний - трансляція (переважно) **Специфічність одного типу клітин:** Збагачена група (ранні сперматиди, олігодендроцити, гальмівні нейрони, пізні сперматиди, збуджуючі нейрони, клітини-попередники олігодендроцитів, клітини мікроглії, астроцити) **Специфічність клітинної лінії:** Cancer enhanced (Rhabdoid) **Тканинна специфічність:** Низька тканинна специфічність	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер раку нирки (сприятливий) та уротеліального раку (несприятливий)
LPAAQVQAF	ENSG00000166704.11 (ZNF606)	9 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічна – невідома функція (переважно) **Специфічність клітинної лінії:** Низька специфічність раку	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер уротеліального раку (сприятливий)

ДОДАТОК 5(Опис неоепітопів із зразка TCGAA2A0SX01A12RA08407)

Послідовність неоепітопу	Білок-кодуючий ген походження	Варіанти сплайсингу	Експресований у	Зв’язок з раком

(kmer, k=9)				
SAAGKEQRV	ENSG00000160256.13 (C21orf70)	5 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Мієлоїдний лейкоз - транспорт кисню (переважно) **Специфічність клітинної лінії**: Низька специфічність раку **Одноклітинний тип кластер експресії:** Неспецифічний - зв'язування РНК (переважно)	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер при раку печінки (несприятливий) і раку нирки (несприятливий)
DLLDEEEGS	ENSG00000165689.17 (ENTR1)	11 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** Неспецифічна – невідома функція (переважно) **Тканинний профіль:** загальна цитоплазматична експресія.(переважно) **Специфічність клітинної лінії:** Низька специфічність раку	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Прогностичний маркер раку нирки (несприятливий)
FLLQRSDTS	ENSG00000102384.13 (CENPI)	8 транскрипти (сплайс-варіанти)	**Лінія клітин кластер експресії:** **Неспецифічний - зв'язування РНК (переважно)** **Тканинний профіль:** загальна цитоплазматична експресія(переважно) **Специфічність клітинної лінії:** Низька специфічність раку	**Специфічний для раку:** Низька специфічність раку **Прогнозне резюме:** Генний продукт не є прогностичним