КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
ННЦ «Інститут біології та медицини»
Кафедра фізіології людини і тварин

Завідувач кафедри: доктор біологічних наук, професор Макарчук М.Ю.
Протокол №____ засідання кафедри
від “____” _____2023 р.

ПОРІВНЯЛЬНИЙ АНАЛІЗ ДИФЕРЕНЦІЙНОЇ ЕКСПРЕСІЇ ГЕНІВ ПРИ ПАТОГЕНЕЗІ НЕЙРОДЕГЕНЕРАТИВНИХ ЗАХВОРЮВАНЬ ТА ДОСЛІДЖЕННЯ КОРЕЛЯЦІЙ КЛЮЧОВИХ ФАКТОРІВ ДЕГЕНЕРАЦІЇ

Кваліфікаційна робота магістра
денної форми навчання
за спеціальністю 091 «Біологія»
Корнєєвої Єлизавети Климентівни

Науковий керівник від кафедри
доктор біологічних наук, професор
Макарчук М.Ю.

Робота виконана на базі ННЦ «Інституту біології та медицини»
під керівництвом д.б.н., професора Макарчука М.Ю.

Оцінка захисту роботи:
_____

Київ – 2023 р.

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

БАС – бічний аміотрофічний склероз;

ГНП – гексануклеотидний повтор;

ДЕГ – диференційна експресія генів;

НДЗ – нейродегенеративні захворювання;

РЗБ – РНК-зв’язуючий білок;

РНК-сек – РНК-секвенування;

ФДТ – фронтотемпоральна деменція (лобно-скронева деменція);

ХА – хвороба Альцгеймера;

ЦНС – центральна нервова система;

C9ORF72 – C9 Open Reading Frame 72 (ген 72 відкритої рамки зчитування хромосоми 9);

CCAR2 – cell cycle and apoptosis regulator 2 (регулятор клітинного циклу та апопотозу 2);

GFF3 – Gene Feature Format version 3 (формат характеристики гену версії 3);

HNRNP – heterogeneous nuclear ribonucleoproteins (гетерогенні ядерні рибонуклеопротеїни);

PCA – principle component analysis (аналіз основних компонент);

STAR – Spliced Transcripts Alignment to a Reference (вирівнювання об’єднаних транскриптів за референсом);

TDP-43 – TAR DNA-binding protein 43.

ЗМІСТ

ВСТУП

Якість існування сучасного населення з істотним подовженням середньої тривалості життя значною мірою визначається нормальним перебігом процесу старіння нейронів центральної нервової системи (ЦНС) і, особливо, опосередковано пов’язана із виникненням захворювань, що характеризуються прискореною втратою нейронів: захворюваннями, які традиційно називаються нейродегенеративними (НДЗ). НДЗ являють собою головну причину захворюваності та інвалідності літніх людей та привертають все більшу увагу дослідників, оскільки формують значний соціально-економічний вплив, що частково пов’язано із загальною тенденцією старіння суспільства (Marcos-Rabal *et al.*, 2021). Ушкодження нейронів є патологічною ознакою хвороб Альцгеймера та Паркінсона, бічного аміотрофічного склерозу (БАС), хвороби Гентінгтона, лобно-скроневої (фронтотемпоральної) деменції (ФТД), розсіяного склерозу та багатьох інших захворювань, які є поширеними серед усіх країн світу та охоплюють значну частку населення (Stephenson *et al.*, 2018).

Дослідження патогенезу та лікування цих розладів стрімко розвиваються. Значною мірою цьому сприяли досягнення в фізіології та генетико-молекулярній біології, оскільки зараз з’являється все більше даних про мутації та відхилення, що корелюють із патогенезом НДЗ. Незважаючи на актуальність терапевтичних втручань, частою закономірністю є недостатній рівень вивчення НДЗ на базовому молекулярному рівні. Недосконало вивчені особливості та відкритті питання в дослідженнях НДЗ створюють багато потенційних перспектив щодо покращення стратегій лікування даних розладів (Conlon and Manley, 2017).

Порушення в будові та функціях РНК-зв’язуючих білків (РЗБ) і дефекти процесів обробки РНК все частіше визнаються як критичні детермінанти неврологічних захворювань. РЗБ рідко функціонують самостійно, а частіше формують взаємодії за сценаріями білкок-білкок та білок-РНК у великій кількості варіацій, що дозволяють здійснювати контроль експресії генів за

певних умов (Nussbacher *et al.*, 2019). Одними із представників РЗБ, що дедалі частіше розглядаються в контексті НДЗ, являються гетерогенні ядерні рибонуклеопротеїни (HNRNP), яким притаманні різноманітні та багатофункціональні ролі в усіх аспектах процесингу мРНК. Роль цих білків у нейродегенерації не є досконало зрозумілою та на даний момент активно вивчається дослідниками. Іншим модулятором, що активно досліджується при НДЗ є TDP-43 (TAR DNA-binding protein 43), який має властивість як РЗБ, так і ДНК-зв'язуючого білка. TDP-43 грає важливу роль у функціонуванні ЦНС. В деяких НДЗ спостерігається накопичення або агрегація TDP-43 в цитоплазмі клітин нервової системи. Це призводить до різних патологічних наслідків, включаючи зниження ефективності транскрипції, порушення обробки мРНК, збільшення експресії генів, що сприяють нейродегенерації, та зменшення стабільності синаптичних з'єднань. Наразі механізми, що призводять до накопичення TDP-43 в цитоплазмі клітин нервової системи, ще не повністю зрозумілі. Проте, деякі дослідження вказують на роль мутацій у генах, що кодують функції метаболізму РНК, як однієї з можливих причин накопичення TDP-43 в цитоплазмі. У дослідженні НДЗ аналіз диференційної експресії генів (ДЕГ) є важливим і одним із центральних інструментів для виявлення ключових генів, які можуть бути пов'язані з розвитком захворювання, а також – є корисним інструментом при дослідженні різних фенотипів захворювання.

В даній роботі було проведено біоінформатичний аналіз та досліджено особливості ДЕГ при патологіях, що часто супроводжують патогенез НДЗ – зразки мозку пацієнтів із накопиченими агрегатами нерозчинного білку HNRNPH1 при нявності гексануклеотидних повторів (ГНП) в гені C9ORF27 (C9) та зразки клітинної лінії, що експресує мутантні форми S375E і S375G білку TDP-43. Було детально досліджено роль генів зі спільним напрямком регуляції експресії в патогенезі НДЗ та їх потенційне значення для діагностики та лікування НДЗ.

РОЗДІЛ 1

ОСОБЛИВОСТІ НЕЙРОДЕГЕНЕРАТИВНИХ ЗАХВОРЮВАНЬ ТА ЇХ ВЗАЄМОЗВ'ЗОК ІЗ ПАТОЛОГІЯМИ TDP-43 та HNRNPH1

1.1. Загальна характеристика та класифікація нейродегенеративних захворювань

Нейродегенеративні захворювання (НДЗ) являють собою визначальну проблему у сфері охорони здоров'я. Зростання клінічних випадків НДЗ, як не парадоксально, початково пов'язують із досягненнями та розвитком медичної сфери, що призвели до зростання середньої тривалості життя населення світу. Між 2000 і 2050 роками частка населення у віці старше 60 років подвоїться приблизно з 11% до 22% (Cummings and Miller, 2007). Зі збільшенням віку ймовірність розвитку НДЗ також зростає. Усі НДЗ поступово прогресують, призводять до інвалідності та є серйозною загрозою для якості життя, цілісності населення та економічних ресурсів.

Нейродегенеративні розлади характеризуються прогресуючою втратою селективно вразливих популяцій нейронів, що контрастує зі статичною втратою нейронів через метаболічні або токсичні дефекти. НДЗ можна класифікувати за первинними клінічними ознаками (наприклад, деменція, хвороба Паркінсона або захворювання рухових нейронів), за анатомічним розподілом нейродегенерації (наприклад, лобно-скроневі дегенерації, екстрапірамідні розлади або спинномозочкові дегенерації) або ж - за основними молекулярними аномаліями (Dugger and Dickson, 2017). Найпоширенішими категоріями НДЗ є амілоїдози, тауопатії, синуклеінопатії та протеїнопатії TDP-43. Зростаюча кількість експериментальних даних свідчить про те, що аномальні білкові конформери можуть поширюватися від клітини до клітини по анатомічно пов'язаних шляхах, що може частково пояснити специфічні анатомічні закономірності, які спостерігаються під час розтину.

НДЗ також класифікуються за їх клінічними проявами. При такій класифікації екстрапірамідні/пірамідні рухові розлади та когнітивні/ поведінкові розлади є найбільш поширеними. Для НДЗ є характерною змішаність клінічних ознак та іноді перебіг декількох супутніх захворювань. Хоча НДЗ, як правило, визначаються специфічним накопиченням білка та анатомічними модифікаціями, - їм також властиві фундаментальні процеси, що пов'язані з прогресуючою дисфункцією і смертю нейронів: протеотоксичний стрес та патології процесу убіквітинування, які регулюються через протеасомну та аутофагосомну/лізосомну системи, окислювальний стрес, апоптоз та нейрозапалення (Dugger and Dickson, 2017). Аномалії білка та гена, що його кодує, які визначають НДЗ, можуть бути присутніми і до появи перших клінічних симптомів (Kanaumi *et al.*, 2013).

На даний момент діагностичні біомаркери НДЗ є досить невизначеними, за винятком рідкісних випадків, коли може бути доведено, що певна генетична мутація викликає розлад (Gómez-Tortosa *et al.*, 2017). Отже, специфічні біомаркери, включаючи біорідинні та молекулярні маркери візуалізації, є основним пріоритетом сучасних досліджень НДЗ.

Найпоширенішими нейродегенеративними розладами є амілоїдози, тауопатії, α-синуклеінопатії та протеїнопатії РНК/ДНК-зв'язуючого білка 43 (TDP-43). Аномальні конформації білка при цих порушеннях та їх клітинний та нейроанатомічний розподіли становлять основні гістопатологічні ознаки, необхідні для постановки конкретного діагнозу. До прикладів накопичень білків в нейронах відносяться: тау в нейрофібрилярних клубках або тільцях Піка, а-синуклеїн в тільцях Леві і TDP-43 у нейрональних цитоплазматичних і нейрональних внутрішньоядерних включеннях. Ці аномальні білкові агрегати складаються з внутрішніх нейрональних білків та інших клітинних компонентів, які виявляються при вірусних інфекціях, де білок є чужорідним. У багатьох випадках білок має аномальну конформацію з амілоїдними властивостями (Dugger and Dickson, 2017).

Перехресні патоморфологічні оцінки стану головного мозку пацієнтів із різними захворюваннями показали, що багато нейродегенеративних розладів мають стереотипне прогресування, яке можна описати за стадіями. Схеми стадій розвитку були визначені для хвороби Альцгеймера та Паркінсона, деменції з тільцями Леві, БАС, ФТД, пов'язаної з патологією TDP-43, і хронічної травматичної енцефалопатії (Coughlin *et al.*, 2019).

1.2. Взаємозв’язок РНК-зв’язуючих білків та нейродегенеративних захворювань

Білки, що зв’язують РНК, мають вирішальне значення для підтримки транскриптому через контрольовану регуляцію процесингу та транспорту РНК. Зміни в роботі та / або структурі цих білків впливають на етапи життєвого циклу РНК, що призводить прояву до різних молекулярних фенотипів. Порушення процесу зв’язування РНК білками і дефекти обробки РНК все частіше визнаються як критичні детермінанти неврологічних захворювань (Nussbacher *et al.*, 2019).

Протягом свого життєвого циклу інформаційні РНК (мРНК) проходять широкомасштабні етапи обробки, включаючи сплайсинг, поліаденілування, редагування, транспортування, трансляцію тощо. Цей складний процес є дуже динамічним і вимагає складної взаємодії між РНК-зв’язуючими білками (РЗБ) для точної модуляції спільної та пост-транскрипційної обробки транскриптів. РЗБ зв’язують молекули РНК у певних послідовностях або вторинних структурах, щоб полегшити етапи процесингу РНК, як у ядрі, так і в цитоплазмі. Геномні дослідження надали основну інформацію про те, як РЗБ впливають на перебіг багатьох фізіологічних процесів на молекулярному рівні (Kim, Kim and Lee, 2021). Дійсно, РЗБ все частіше визнаються як багатофункціональні білки, оскільки специфічний РЗБ може асоціюватися з

різними білковими комплексами, щоб впливати на декілька етапів обробки його РНК-мішеней. Тому не дивно, що змінена взаємодія між РЗБ та його мішенями може призвести до важких патологічних фенотипів. Фактично, результати багатьох досліджень підтверджують, що дефекти регуляції РНК відіграють вирішальну роль у дисфункції нервового розвитку та НДЗ. Механізми, під впливом яких функціонування РЗБ змінюються при неврологічних розладах, є дуже різноманітними, і в деяких випадках молекулярні деталі залишаються на даний момент невідомими. Однак, вже було встановлено певні кореляції між функціональністю певних РЗБ та НЗД, що включають порушення експресії РЗБ, неправильну клітинну локалізацію та агрегацію РЗБ, а також - секвестрацію РЗБ транскриптами та / або аномальними білками з патологічними повторними розширеннями (рис. 1.2).

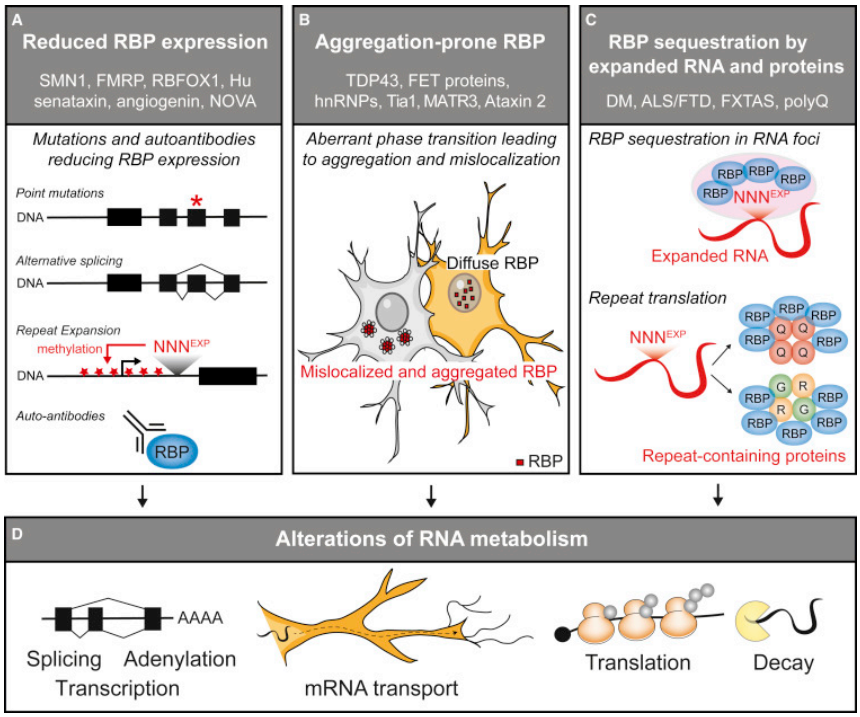

Рис. 1.1. Механізми, що призводять до порушення роботи РЗБ при нейродегенеративних захворюваннях (Nussbacher *et al.*, 2019).

Об’єднуючою рисою усіх РЗБ, пов’язаних із нейродегенеративними захворюваннями, є наявність пріоноподібних доменів, які сприяють утворенню агрегатів (Nussbacher *et al.*, 2019). Пріоноподібні домени

характеризуються послідовностями низької складності, а не конкретною первинною послідовністю, і збагачені полярними незарядженими амінокислотами, такими як глутамін, аспарагін, тирозин, серин і гліцин (Wang *et al.*, 2018). Такі домени низької складності, що складаються лише з кількох амінокислот, були передбачені у понад 200 білках у геномі людини із суттєвою перевагою групи білків, що зв'язують РНК та ДНК. Ці домени мають вирішальне значення для динамічного складання та розбирання рибонуклеопротеїнових гранул, а також, - надають властивість агрегації декільком РЗБ, пов'язаним із НДЗ (Prashad and Gopal, 2021).

Визнання РЗБ як основного фактора розвитку НДЗ ілюструється конкретними прикладами та описами загальногеномних досліджень, які підтверджують вплив поширеної неправильної регуляції обробки РНК на патогенез захворювання. РНК-зв'язуючі білки відіграють центральну роль у регулюванні всіх аспектів експресії генів, тому їхня дисфункція, ймовірно, є ключовою ознакою порушення гомеостазу РНК та білків при НДЗ. З'являються методи терапевтичного відновлення функцій РЗБ, а також обговорюються поточні стратегії боротьби з порушеннями гомеостазу РЗБ (Nussbacher *et al.*, 2019).

1.3. Особливості функцій білків HNRNP

Родина гетерогенних ядерних рибонуклеопротеїнів (HNRNP) — це родина РЗБ, що містять один або декілька РНК-зв'язуючих доменів, які сприяють їх широкій і дивергентній функціональності на всіх стадіях метаболізму нуклеїнових кислот (Geuens, Bouhy and Timmerman, 2016).

Домени, які включають найбільш закономірний мотив розпізнавання РНК, домен гомології K і мотив RGG, надають HNRNP здатність зв'язувати велику кількість РНК-мішеней у межах обширного інтерактому. Примітно, що, як і інші РЗБ, HNRNP також можуть зв'язувати РНК через свої внутрішньо

невпорядковані області або домени низької складності, як їх частіше називають. Це області низької амінокислотної складності, що сприяють утворенню рибонуклеопротеїнових комплексів вищого порядку за допомогою фазового поділу, керованого доменами низької складності (Hofweber and Dormann, 2019). Хоча HNRNP можуть взаємодіяти з партнерами, що зв'язують РНК, специфічним для послідовності чином, неспецифічні взаємодії також відомі серед HNRNP, що узгоджується з їх різноманітними функціональними можливостями. Деілька представників родини HNRNP також містять послідовності ядерної локалізації та здатні модулюватись ядерними експортними сигналами, які опосередковують їх переміщення до та з цитоплазми.

Функціонально білки HNRNP задіяні на всіх етапах дозрівання мРНК, включаючи регуляцію транскрипції, кепінг, альтернативний сплайсинг, поліаденілування, транспорт і підтримку стабільності. Локалізація HNRNP відбувається переважно в ядрі, однак, декілька HNRNP можуть пересуватись між ядром і цитоплазмою для забезпечення регулювання додаткових цитоплазматичних функцій (Michael, Eder and Dreyfuss, 1997). Дійсно, HNRNP утворюють високодинамічні комплекси з РНК та іншими РЗБ для регуляції цих процесів. Вони здатні успішно асоціюватися та взаємодіяти з безліччю різних механізмів обробки мРНК завдяки постійному ремоделюванню їх складних комплексів мРНК-білок.

Велика кількість представників HNRNP прямо чи опосередковано причетні до патогенезу НДЗ. Це не дивно, враховуючи велику кількість взаємодій родини HNRNP як один з одним, так і з ключовими генами та білками, пов'язаними з НДЗ, включаючи TDP-43, C9ORF72, FUS і Tau (Bampton *et al.*, 2020). Зростає кількість доказів, які також свідчать про те, що інші білки HNRNP залучаються не тільки до класичних включень, що спостерігаються при НДЗ, але й до патологій, пов'язаних з мутацією розширення гену C9ORF72 за рахунок гексануклеотидних повторів (ГНП) (Bampton *et al.*, 2020).

TDP-43 і FUS, ймовірно, є найбільш відомими HNRNP в області нейродегенерації. Їх накопичення в патологічних включеннях при БАС і ФТД лежить в основі сучасних досліджень механізмів даних захворювань. Аномальне відкладення TDP-43 є основною нейропатологічною ознакою в 97% випадків БАС і ~ 50% випадків ФТД (ФТД-TDP), і тому БАС та ФТД часто групуються разом як протеїнопатії TDP-43 (Scaber and Talbot, 2016).

1.4. TDP-43 протеїнопатії

TDP-43 являє собою РНК/ДНК-зв'язуючий білок розміром 43 кДа, який відіграє роль у репресії транскрипції, модуляції альтернативного сплайсингу та метаболізмі РНК. TDP-43 кодується геном TARDBP і часто класифікується як член сімейства HNRNP, але не називається таким через те, що його пропустили на початкових експериментах з двовимірним гелем та імуноочищенням (Dreyfuss, Kim and Kataoka, 2002).

У нормальних клітинах TDP-43 переважно присутній в ядрі і відіграє важливу роль у регуляції РНК, наприклад, у регуляції транскрипції, альтернативному сплайсингу та стабілізації мРНК. За патологічних умов може відбуватися розщеплення, гіперфосфорилювання та убіквітинування TDP-43 (Jo *et al.*, 2020). Накопичення агрегатів TDP-43 в ЦНС є загальною ознакою багатьох нейродегенеративних захворювань, таких як бічний аміотрофічний склероз (БАС), фронто-темпоральна деменція (ФТД), хвороба Альцгеймера (ХА) та лімбічна переважаюча вікова TDP-43 енцефалопатія (Jo *et al.*, 2020). Пацієнти з ХА та патологією TDP-43 мають підвищену вираженість когнітивних порушень порівняно з пацієнтами без патології TDP-43. Крім того, найпоширеніший генетичний фактор ризику ХА, аполіпопротеїн Е4 (APOE4), асоціюється з підвищеною частотою патології TDP-43 (Meneses *et al.*, 2021).

У 2006 році було виявлено, що TDP-43 є основним компонентом нейрональних включень при БАС та ФТД із включеннями убіквітину, які тепер називаються ФТД-TDP (Bennion Callister and Pickering-Brown, 2014). TDP-43 зазвичай є ядерним білком, але при НДЗ він утворює включення в цитоплазмі, ядрі та клітинних відростках (рис. 1.2).

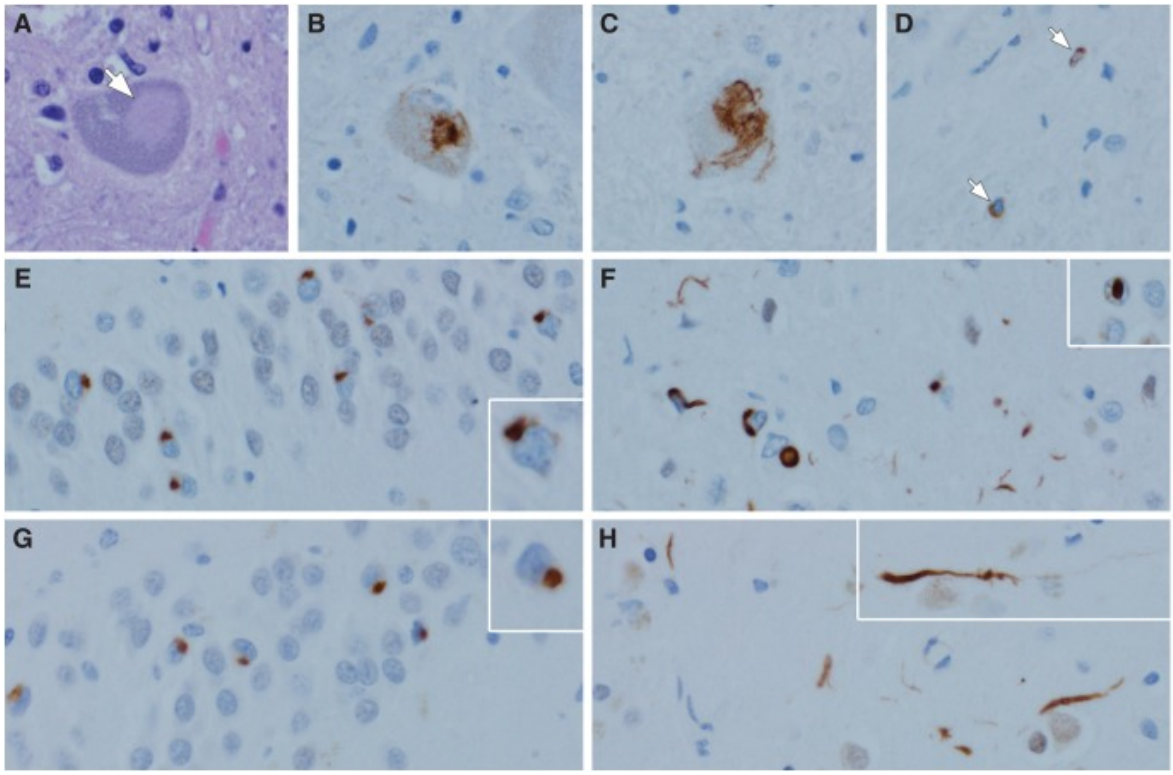

Рис. 1.2. Включення TDP-43 при: БАС (A–D) і ФТД з нейрональними включеннями TDP-43 (ФТД-TDP) (F–H) (Dugger and Dickson, 2017).

Початкові докази, що вказують на роль процесингу РНК при БАС і ФТД, були новаторським відкриттям TDP-43 як компонента цитоплазматичних і убіквітинованих включень у нейронах пацієнтів зі спорадичним БАС та ФТД (Arai *et al.*, 2006). За цим висновком швидко послідувало визначення мутацій TDP-43 як генетичних причин БАС та ФТД (Lagier-Tourenne, Polymenidou and Cleveland, 2010). При БАС ядерне очищення TDP-43 супроводжується накопиченням білка у цитоплазматичних включеннях. Навпаки, картина відкладення TDP-43 у патологічному спектрі ФТД-TDP є набагато

гетерогеннішою з різноманітними морфологічно відмінними цитоплазматичними та внутрішньоядерними імунореактивними включеннями TDP-43, що характеризують п’ять молекулярних підтипів захворювання. Було показано, що HNRNP E2 спільно локалізується у включеннях типу C ФТД-TDP, пов’язаних із семантичною деменцією, а нещодавно – у включеннях типу A (Bampton *et al.*, 2020). Виснаження TDP-43 призводить до дестабілізації цільових транскриптів, які є вираженими в тканинах мозку пацієнтів при ФТД та БАС (Bampton *et al.*, 2020).

Окрім БАС та ФТД-TDP, аномальний TDP-43 можна виявити в 25-50% випадків ХА, переважно в лімбічному розподілі (Kansal *et al.*, 2015). Крім того, у більшості випадків склерозу гіпокампу у літніх людей, який характеризується селективною втратою нейронів, TDP-43 впливає на сектор CA1 гіпокампу та пов’язаний із перебігом амнестичного клінічного синдрому (Dugger and Dickson, 2017).

Механізм та причини агрегації білкових включень та інших фізіологічних порушень TDP-43 при НДЗ не є досконало зрозумілими. Однією з причин такої патології вважається мутація окремих ділянок білка. Нещодавно було описано мутацію в TDP-43 у випадку раннього початку БАС, яка впливала на потенційний сайт фосфорилювання в положенні 375 (S375G). Було продемонстровано, що як мутація S375G, так і її фосфоміметичний варіант, S375E, демонструють змінений ядерно-цитоплазматичний розподіл та клітинну токсичність (Paron *et al.*, 2022). В дослідженні (Paron *et al.*, 2022) було виявлено, що ці мутанти, схоже, не впливають на добре вивчені аспекти TDP-43, такі як сплайсинг РНК, авторегуляція, конформація, динаміка чи агрегація білка, але демонструють дисморфну форму ядер і зміни в клітинному циклі. Просеквеновані данні дикого типу (контролю) та мутантних форм TDP-43 S375E і S375G було взято за основу досліджень цієї магістерської роботи.

1.5. Патології C9ORF72 та їх взаємозв'язок із HNRNPH1

Надійні докази приналежності ФТД і БАС до спектру захворювань були отримані з відкриття, що розширення гексануклеотидних повторів (ГНП) у першому інтроні гена 72 відкритої рамки зчитування хромосоми 9 (C9ORF72) є найбільш поширеною генетичною причиною спадкових форм ФТД (C9-FTD) і БАС (C9-ALS) або разом C9-FTD/ALS (Renton *et al.*, 2011). Фенотипово включення TDP-43 асоціюються з більшістю цих випадків, і багато носіїв ГНП відповідають клініко-патологічним діагностичним критеріям для обох розладів. Було запропоновано та розглянуто два C9ORF72 ГНП-опосередкованих токсичних посилення функціональних механізмів: а саме - токсичність РНК, опосередкована внутрішньоядерними вогнищами РНК та включення білка дипептидного повтору з неканонічно трансльованих транскриптів ГНП (Balendra and Isaacs, 2018). Крім того, втрата функції білка C9ORF72 була запропонована як патогенний механізм, однак, хоча було показано, що знижена функція C9ORF72 посилює механізми токсичності, втрата білка є недостатньою для відтворення фенотипу захворювання у ссавців.

Вважається, що вогнища РНК (результат розширених повторів РНК) виявляють свою токсичність шляхом секвестрування, викликаючи функціональну втрату ключових РЗБ. Дослідження показали, що ізоформи HNRNPH1 і HNRNPH3 є найбільш послідовно знайденими білками, які асоціюються з ГНП в клітинних і тваринних моделях (Haeusler *et al.*, 2014). Подальше дослідження того, як рівні нерозчинних агрегатів HNRNP можуть модулювати індуковану токсичність у нейронах, може пояснити патомеханічну основу їхнього залучення до патогенезу НДЗ (Bampton *et al.*, 2020).

РОЗДІЛ 2

МАТЕРІАЛИ ТА МЕТОДИ ДОСЛІДЖЕНЬ

2.1. Програмне забезпечення та пакети

2.1.1. Використані бази даних

Для даної роботи були використані такі бази даних:

1. NCBI Gene Expression Omnibus (GEO) - це міжнародний архів наборів даних експресії генів та елементів функціональної геноміки, що був започаткований Національним центром біотехнологічної інформації (NCBI - National Center for Biotechnology Information). В NCBI GEO зберігаються архіви даних РНК секвенування (РНК-сек), які номенклатурно структуровані за допомогою системи індивідуальної індексації для кожного окремого набору даних (Barrett *et al.*, 2013). Базу даних NCBI GEO було використано для завантаження послідовностей РНК-сек експерименту TDP-43 у форматі FASTQ для подальшого аналізу.

2. GENCODE – це геномний проект, створений за сприяння консорціуму ENCODE, та направлений на дослідження і класифікацію геномних характеристик організму людини та миші (Frankish *et al.*, 2021). База даних GENCODE була використана для завантаження анотації геному людини.

3. ENSEMBL – це проект бази даних геномів, що слугує централізованою платформою для дослідників, які вивчають геноми різних хребетних, модельних організмів та людини (Cunningham *et al.*, 2022). В даній роботі цей ресурс було використано для анотації генів при виконанні аналізу диференційної експресії генів.

2.1.2. Програмне забезпечення, сервери та інструменти біоінформатичного аналізу

Для даної роботи були використані наступне програмне забезпечення, сервери та інструменти біоінформатичного аналізу для обробки даних:

1. Сервер Galaxy (домени ORG та EU) був використаний для обчислювальних операцій на першоетапній стадії обробки даних секвенування (Afgan *et al.*, 2018).

2. Інструмент Fasterq-dump (версія 3.0.3+galaxy0;) був використаний для завантаження та екстракції даних секвенування у форматі fastq із архіву NCBI GEO.

3. Інструмент FastQC (версія 0.73+galaxy0) був використаний для перевірки та контролю якості проміжних результатів біоінформатичної обробки даних, зокрема, - після застосування інструменту Fasterq-dump.

4. Інструмент MultiQC (версія 1.11+galaxy1) був використаний для об'єднання результатів біоінформатичного аналізу в єдиний звіт, зокрема, - після застосування інструменту FastQC до завантажених даних у FASTQ форматі, а також – після проведення геномного вирювнювання послідовностей (див. пункт 5) для файлів формату bam.

5. Інструмент STAR (версія 2.7.10b+galaxy3) був використаний для проведенення геномного вирювнявання завантажених даних із послідовністю референсного геному.

6. Інструмент htseq-count (версія 0.9.1+galaxy1) був використаний для підрахунку вирівняних рідів у форматі bam відповідно до кожного із генів референсного геному.

2.1.3. Аналіз диференційної експресії генів та візуалізація результатів

Для дослідження диференційної експресії генів, кореляційного аналізу та візуалізації результатів були використані наступні програмні пакети та інструменти RStudio (версія 4.0.3):

1. DESeq2 (версія 1.30.1) – даний пакет використовувався для оцінки варіативної залежності даних секвенування (дані підрахунку рідів read counts)

та дослідження диференційної експресії генів на основі моделі негативного біноміального розподілу.

2. pcaExplorer (версія 2.16.0) – даний інструмент використовувався для аналізу головних компонент даних (PCA – principle component analysis) та їх кластерного розподілу.

3. biomaRt (версія 2.46.3) – даний пакет надає можливість отримувати доступ до великих обсягів даних в уніфікований спосіб без необхідності застосування SQL-запитів. В даній роботі пакет biomaRt було використано для встановлення доступу до бази даних Ensembl.

Програмні пакети для візуалізації кроків аналізу та отриманих результатів:

1. ComplexHeatmap (версія 2.6.2) та pheatmap (версія 1.0.12) – дані пакети було використано для створення так званих «теплових» карт подібності з метою порівняння експериментальних даних і визначення їхньої подібності.

2. RColorBrewer (версія 1.1-3) - при використанні R пакет RColorBrewer є важливим інструментом для керування кольором, що надає можливість використання різноманітних палітр і слугує корисним інструментом для візуалізації.

3. ggplot2 (версія 3.3.5) – даний пакет призначений для створення складних графіків на основі даних, що зберігаються у табличній формі. Він представляє програмний інтерфейс, який дозволяє користувачам вказувати, які змінні відображати на графіку, як вони мають бути візуалізовані та інші характеристики. Цей пакет було використано для багатоетапної координатної візуалізації диференційної експресії генів.

4. limma (версія 3.46.0) та ggvenn (версія 0.1.9) – дані пакети було використано для підрахунку кореляції диференційної експресії генів та відображення результатів у вигляді діаграм Венна.

5. magrittr (версія 2.0.3), ggrastr (версія 1.0.1), dplyr (версія 1.0.8) – інші інтегровані пакети R, які було використано для забезпечення базових можливостей роботи із даними.

2.2. Експериментальні дані

В даній роботі досліджувалися набори даних, що пов'язані з одними із найбільш закономірних патологічних особливостей нейродегенеративних захворювань. До них відносяться:

(1) Мутаційно-опосередковані патології TDP-43, які при нейродегенеративних розладах часто супроводжуються утворенням убіквітин-позитивних і тау-негативних включень.

(2) Скупчення нерозчинних агрегатів, які також розповсюджено являють собою РНК-зв'язуючі білки (РЗБ) і виступають в ролі регуляторів альтернативного сплайсингу. В даному контексті було розглянуто набори даних пацієнтів із високими рівнями нерозчинних агрегатів білку HNRNPH1, для яких також була характерною наявність гексануклеотидних повторів у гені C9ORF72 (C9).

2.2.1. Дані мутантних форм TDP-43

В цій роботі було використано дані, що були опубліковані 2022 року в статті 'Alterations in cell cycle progression caused by S375G and S375E phosphomutants of TDP-43' (Paron *et al.*, 2022). Для вивчення ефектів мутацій сайту фосфорилювання в положенні 375 – S375G та S375E – вченими було створено клітинні лінії, що експресують дикий тип TDP-43 (контроль), а також

- варіанти S375G та S375E. Дизайн клітинної лінії HEK 293 Tet-On 3G являє собою трансформовану клітинну лінію ембріональних нирок людини, яка експресує тетрациклін (Tet)-регульований трансактиватор Tet-On 3G. Клітинна лінія була згенерована за допомогою системи Flp-In T-Rex – це система, що дозволяє генерувати стабільні клітинні лінії ссавців, які демонструють тетрациклін-індуковану експресію цільового гена. Отримані зразки було просеквеновано за допомогою інструменту Illumina HiSeq 2500.

Для даної магістерської роботи опубліковані дані секвенування було завантажено із бази даних NCBI GEO за ідентифікаторним номером GSE167385 і згруповано в три категорії в залежності від прояву – дикий тип (контроль) та мутантні форми S375G та S375E (табл. 2.1).

Таблиця 2.1

Перелік набору даних з нормальними та мутантними формами TDP-43

Номер зразка	GEO ідентифікатор	SRR ідентифікатор	Прояв
1	GSM5104582	SRR13775756	Дикий тип
2	GSM5104583	SRR13775757	Дикий тип
3	GSM5104584	SRR13775758	Дикий тип
4	GSM5104585	SRR13775759	S375G
5	GSM5104586	SRR13775760	S375G
6	GSM5104587	SRR13775761	S375G
7	GSM5104588	SRR13775762	S375E
8	GSM5104589	SRR13775763	S375E
9	GSM5104590	SRR13775764	S375E

Кожен із зразків складається із даних парнокінцевого (двостороннього) секвенування та містить прямий і зворотній ріди, кожен із яких має довжину 150 нуклеотидів.

2.2.2. Дані пацієнтів із гексануклеотидними повторами в гені C9ORF72 та агрегатами білка HNRNPH1

Другим набором експериментальних даних в цій роботі виступали просеквеновані зразки пацієнтів з підвищеним рівнем нерозчинних агрегатів білка HNRNPH1 та наявністю гексануклеотидних повторів в гені C9ORF72 (C9). Дані були опубліковані у 2018 році та розміщені в архіві NCBI GEO за ідентифікаторним номером GSE116622. Ці дані були також отримані за допомогою парнокінцевого секвенування на платформі Illumina HiSeq 2500. Попереднє опрацювання та біоінформатичний процесинг цих даних були здійснені в рамках моєї бакалаврської роботи за спеціальністю «Біоінформатика» у Франкфуртському університеті ім. Йогана Вольфганга Гете під керівництвом Dr. Kathi Zarnack та Dr. Mario Keller.

Із оброблених даних для цієї магістерської роботи було відокремлено 3 зразки з групи C9 із високим рівнем нерозчинного HNRNPH1 (C9-high). До магістерської роботи було перенесено файли з результатами геномного вирівнювання у форматі bam, які попередньо були перевірені за параметрами якості, з метою подальшого проведення аналізу диференційної експресії генів та порівняння даних зразків із патологією в гені TDP-43.

2.3. Попередня обробка даних TDP-43

Усі етапи попередньої обробки даних здійснювались на біоінформатичному кластері Galaxy доменних версій EU (дані TDP-43) та ORG (дані C9-high).

Для даних патологій TDP-43 було сформовано перелік ідентифікаторних номерів SRR архіву SRA, за якими зразки можливо було відрізнити між собою в межах загального експерименту, що був розміщений в базі даних NCBI GEO. Це було зроблено за допомогою інтегрованого інструменту, який

пропонується архівом NCBI GEO, - SRA Run Selector. Таким чином, ідентифікаторні номери SRA кожного із зразків було скопійовано до серверу Galaxy, після чого до них було застосовано інструмент fasterq-dump, який завантажує до серверу усі зразки, індентифікатори SRA яких представлені у списку, у форматі FASTQSANGER.GZ. Інструмент fasterq-dump автоматично розпізнає тип секвенування зразків, який було застосовано, в цьому випадку – парнокінцеве. В результаті застосування даного інструменту було сформовано картотеку експериментів, яка складалась із 9 файлів просеквенованих послідовностей (3 зразки з мутацією S375G, 3 зразки з мутацією S375E та 3 контрольні зразки).

З метою перевірки якості просеквенованих даних, до кожного із зразків було застосовано інструмент FastQC, який дозволяє оцінити рівень якості даних при їх обробці, зокрема, відображає базову статистичну інформацію, показники якості вздовж усіх просеквенованих нуклеотидів та послідовностей, а також - відсоткові вмісти нуклеотидів GC, кількість дуплікатів та залишкових адапторних послідовностей. Кожен із зразків мав два результативні звіти інструменту FastQC, окільки перевірка якості в даному випадку застосовується для кожного з рідів окремо (прямого та зворотнього). Звіти інстременту FastQC для кожної із послідовностей було згруповано за допомогою інструменту MultiQC, який дозволяє формування загального підсумку множинних файлів з метою полегшення візуалізації результатів. Таким чином, кожен окремий звіт FastQC було відтворено в загальному результативному підсумку, який було згенеровано за допомогою MultiQC.

Наступним кроком було проведення геномного вирівнювання послідовностей. Геномне вирівнювання послідовностей - це процес порівняння послідовностей ДНК або РНК (в даному випадку РНК) з референсною геномною послідовністю. Цей процес є важливим етапом в аналізі геноміки, оскільки дозволяє встановити місцеположення генів та визначити варіанти змін в геномі. Інструмент STAR (Spliced Transcripts Alignment to a Reference) - це програмне забезпечення, що використовується

для геномного вирівнювання РНК-послідовностей (РНК-сек) з референсним геномом. Основна перевага інструменту STAR полягає в його здатності вирівнювати РНК-послідовності, які містять інтрони. STAR враховує відкриті рамки зчитування та підтримує детальне розмежування екзонів та інтронів. STAR використовує алгоритми, які забезпечують високу швидкість та точність вирівнювання, в тому числі враховуючи мультиплексність та корекцію помилок. Після вирівнювання STAR генерує вихідні файли, які містять інформацію про вирівняні регіони та екзони, кількість прочитань в кожному гені та іншу статистичну інформацію, яка може бути використана для подальшого аналізу.

В даній роботі при проведенні геномного вирівнювання в якості референсного геному був використаний геном людини, завантажений із бази даних GENCODE, випуску 39 (GRCh38.p13) у форматі GFF3. GFF3 (Gene Feature Format version 3) - це текстовий формат для представлення геномних або транскриптомних анотацій, що містить інформацію про геномні функціональні елементи, такі як гени, екзони, інтрони, нетранскльовані області, а також - їх атрибути, такі як назви, описи, координати та інші додаткові властивості. Формат GFF3 є стандартом для багатьох біоінформатичних програм та баз даних, таких як NCBI, Ensembl, UCSC Genome Browser та багатьох інших, і дозволяє дослідникам взаємодіяти зі стандартними форматами даних та аналізувати геномні або транскриптомні анотації. Формат GFF3 складається з шести полів, розділених символом табуляції, які описують різні характеристики функціонального елемента:

- Seqid: ідентифікатор хромосоми або сегмента геному.
- Source: джерело, яке згенерувало або забезпечило дані.
- Type: тип функціонального елемента.
- Start: початкова координата функціонального елемента на геномі.
- End: кінцева координата функціонального елемента на геномі.
- Strand: напрямок ріду, на якому знаходиться функціональний елемент.

Окрім вищезазначаених структурних елементів, формат GFF3 також містить додаткові поля, які можуть включати додаткові атрибути та значення для функціональних елементів. Формат GFF3 є розширенням формату GFF (Gene Feature Format) та має більшу кількість полів та удосконалені специфікації, що дозволяють більш детально описати геномні анотації.

Деякі із параметрів проведення геномного вирівнювання встановлювались вручну, враховуючи особливості експерименту та специфіку вимог до процедури вирівнювання (таб. 2.2). Усі інші параметри вирівнювання були використані за встановленими стандартними значеннями, що пропонуються сервером Galaxy.

Таблиця 2.2

Перелік параметрів геномного вирівнювання послідовностей, які було встановлено вручну

Параметр	Обране значення
Одно- чи парнокінцеві послідовності використовуються для вирівнювання	Парнокінцеві (пряма та зворотня послідовності)
Довжина геномної послідовності навколо анотованих з'єднань	149 (що дорівнює довжині послідовності мінус 1)
Максимальна кількість вирівнювань для виведення результатів, плюс 1	1
Максимальна кількість неспівпадінь для виведення вирівнювання, плюс 1	999
Максимальне відношення неспівпадінь до загальної довжини вирівнювання	0.04*

* значення було затверджено до рекомендації на конзорціумі ENCODE

В результаті геномного вирівнювання для кожного із 9-ти експериментальних зразків було виведено статистичні звіти (файли LOG та BED) про проведення вирівнювання, а також – результуючі проанотовані файли у форматі BAM, які містять інформацію про розташування кожного із генів за координатами референсного геному, що дозволяє знайти відповідність між просеквенованою ділянкою та конкретним геномним регіоном. Для

проміжного оцінування якості геномного вирівнювання до звітів у форматі LOG було застосовано інструмент MultiQC, який в даному випадку дозволяє визначити відсоток вирівняних послідовностей від загальної кількості, а також – кількість / відсоток послідовностей, що було вирівняно одразу із декількома геномними ділянками (похибка), що не вдалося вирівняти (занадто короткі послідовності, неспівпадіння або з інших причин).

Завершальним етапом початкової обробки даних став підрахунок кількості прочитань (рідів – reads) за допомогою інструменту ht-seq count. Htseq-count - це інструмент, який дозволяє підрахувати кількість рідів РНК-послідовностей, які припадають на кожен ген у геномі (Srinivasan, Virdee and Mcarthur, 2020). Це робиться на основі вхідних даних у вигляді файлу BAM (Binary Alignment/Map), який містить інформацію про те, які ріди співпадають з кожним геном. Інструмент htseq-count використовує анотаційні файли у форматі GFF або GTF – в даному випадку GFF, – які містять інформацію про місцезнаходження генів у геномі. Htseq-count використовує ці файли для визначення того, які ріди належать до кожного гена. В результаті роботи htseq-count геномна послідовність розбивається на фрагменти, які відповідають окремим генам. Для кожного гена обчислюється кількість рідів, які були прикріплені до нього, тобто їх "покриття". Ця інформація може бути використана для порівняння рівня експресії різних генів у зразках з різною біологічною інформацією, наприклад, з різних тканин або різних умов обробки.

Таким чином, після застосування інструменту htseq-count для зразку даних було створено підрахунки для кожного із просеквенованих генів у табличній формі, після чого ці файли було завантажено на локальний диск та імпортовано до середовища RStudio для проведення аналізу диференційної експресії генів.

2.4. Обробка даних C9ORF72 HNRNPH1

Як згадувалося вище, експериментальні дані патологічних скупчень агрегатів білку HNRNPH1 у пацієнтів із вираженими гексануклеотидними повторами у гені C9ORF72 (C9) було попередньо оброблено із використанням аналогічних методів. Попередня обробка даних була виконана до етапу геномного вирівнювання включно як частина експериментальної роботи до бакалаврського тезису за програмою «Біоінформатика» при Франкфуртському університеті ім. Йогана Вольфганга Гете. Файли у форматі BAM було завантажено із університетського кластеру, де вони зберігались, на локальний жорсткий диск. Після чого, отримані файли із геномним вирівнюванням було імпортовано до серверу Galaxy, де, по аналогії із обробкою даних TDP-43, до даних було застосовано інструмент htseq-count з метою підрахунку кожного із просеквенованих генів. Отримані дані також було завантажено із серверу Galaxy на локальний диск з метою проведення подальшого аналізу диференційної експресії генів у середовищі RStudio та порівняння із патологією TDP-43.

2.5. Аналіз диференційної експресії генів

Диференційна експресія генів (ДЕГ) описує різницю в кількості РНК, що виражається в різних зразках. Теоретично, якщо ген є однаково експресованим в двох зразках, то його РНК повинна бути присутня в однаковій кількості в обох зразках. Якщо ген експресується різним чином в двох зразках, то кількість РНК, що виражається, буде відрізнятися.

Аналіз диференційної експресії генів (DEG) є важливим інструментом у геномній біології та медичній генетиці, оскільки він дозволяє виявити гени, які виявляють зміну в експресії між двома або більше групами зразків

(Stupnikov *et al.*, 2021). Одним із класичних методів, що також залучається в аналізі ДЕГ є принцип аналізу основних компонент. Принцип аналізу основних компонент (PCA) - це статистичний метод, який може допомогти зменшити розмірність даних ДЕГ та відокремити основні (принципові) компоненти варіації в цих даних (Nwakuya and Biu, 2019). У PCA, множина генів розглядається як набір вимірювань у просторі вищих вимірів. Кожен ген можна розглядати як одниницю вимірювання, а різні зразки можна розглядати як різні точки в цьому просторі. PCA шукає головні напрямки варіації в цьому просторі, тобто ті, які пояснюють найбільшу кількість варіації в даних. Зазвичай, головні компоненти обчислюються за допомогою лінійних комбінацій вихідних змінних (генів), які максимізують дисперсію в даних. У контексті ДЕГ, PCA може використовуватися для зменшення розмірності даних, знаходження генів, які мають найбільший внесок у варіацію даних, та для визначення генів, що мають схожу експресію між групами зразків. Після застосування PCA можна використовувати інші методи аналізу, такі як групування та класифікація генів, щоб ідентифікувати потенційні біологічні механізми, які пояснюють ДЕГ між групами зразків.

Для того, щоб оцінити ДЕГ, використовуються різні параметри. Один з таких параметрів - це логарифмічне відношення кількості РНК (Log2Fold change), що виражається в зразках (розбіжність логарифмів) (Li *et al.*, 2022). Цей показник використовується для визначення рівня диференційної експресії генів між зразками. Розбіжність логарифмів обчислюється, порівнюючи кількість РНК, що виражається в досліджуваному зразку, з кількістю РНК, що виражається в контрольному зразку, після чого виконується логарифмування результатів. Цей показник відображає, у скільки разів експресія гену змінюється між двома порівнюваними групами. Зазвичай, значення Log2Fold Change може бути додатнім або від'ємним, в залежності від того, яка з груп має вищу експресію гену. Наприклад, якщо значення Log2Fold Change для певного гену додатнє, це означає, що експресія цього гену в певному зразку вища, ніж в контрольній групі. Якщо ж значення Log2Fold Change від'ємне, це

означає, що експресія гену в контрольній групі вища, ніж в досліджуваному зразку.

Інший важливий параметр, який враховується при обробці даних ДЕГ - це рівень статистичної значущості, який використовується для оцінки ймовірності, що різниця в експресії гену між двома зразками є справжньою (Jiang *et al.*, 2022). Для цього використовуються різні статистичні методи та показники, такі як, наприклад, P-значення (p-value). P-значення є статистичним параметром, який використовується для визначення статистичної значущості різниці в експресії генів між двома або більше зразками при аналізі ДЕГ. P-значення показує ймовірність отримання різниці в експресії генів між зразками, якщо ніякої різниці в експресії не існує (нульова гіпотеза). Чим менше p-значення, тим менш ймовірно, що отримана різниця між зразками може бути пояснена нульовою гіпотезою, тобто тим більша ймовірність, що різниця в експресії генів є статистично значущою. Зазвичай, при аналізі ДЕГ, встановлюють порогове значення для p-значення, яке вказує на максимальне значення p, при якому різниця в експресії генів ще вважається статистично значущою. Як правило, порогове значення для p встановлюється в межах від 0.01-0.05 до 0.1, що вказує на те, що якщо p-значення менше цього значення, то різниця в експресії генів є статистично значущою. Однак, при використанні p-значення необхідно враховувати й інші параметри, такі як кількість зразків та реплікатів, а також, - частоту виявлення генів.

В даній роботі аналіз ДЕГ було виконано в декілька етапів на декількох експериментальних групах. На першому набір даних патології TDP-43 досліджувався індивідуально та обмежувався порівнянням експресії генів в обох мутаційних формах з контролем з метою розуміння динаміки та особливостей мутаційних проявів для випадків S375G та S375E на генному рівні. Після цього зразки мутантних форм TDP-43 було досліджено разом зі зразками високого вмісту нерозчинних агрегатів HNRNPH1 при нявних гексануклеотидних розширеннях в гені C9 (C9-high). Набори даних було

об’єднано для аналізу з метою знаходження відповідностей та/чи розбіжностей в ДЕГ. Доцільність цього корелятивного порівняння була насамперед зумовлена тим, що знаходження певних подібностей в механізмах регуляції генів дозволила б зробити певні висновки та припущення щодо спільних механізмів патогенезу нейродегенеративних захворювань, зокрема таких, для яких є характерними патології TDP-43 та C9 з білковими агрегатами, - лобно-скронева деменція, бічний аміотрофічний склероз, хвороба Паркінсона тощо.

В даній роботі для проведення експериментального аналізу диференційної експресії генів було використано середовище RStudio та програмний пакет DESeq2 в якості основного аналітичного інструменту. Окрім цього, для роботи з геномною інформацією до аналізу було інтегровано доступ до бази даних ENSEMBL, яка дозволила формування анотованої картотеки генів, що були задіяні в аналізі.

РОЗДІЛ 3
РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ ТА ЇХНЄ ОБГОВОРЕННЯ

3.1. TDP-43 біоінформатична обробка та аналіз даних

Набори даних секвенування мутаційних форм TDP-43 S375E та S375G і контрольних груп попередньо оброблялись на сервері Galaxy у форматі колекції даних. Кожен із файлів з отриманими послідовностями пройшли послідовні етапи біоінформатичного аналізу як описувалось в розділі 2.

3.1.1. Перевірка якості просеквенованих даних

Інструмент FastQC було використано для перевірки файлів із просеквенованими послідовностями парних колекцій даних, для кожної з яких було окремо проаналізовано прямий та зворотній ріди. Кожен із рідів для усіх експериментів було згруповано на рівні колекції даних за допомогою інструменту MultiQC, та отримано сумарний звіт по обраним параметрам якості (таб. 3.1).

Таблиця 3.1

Сумарний звіт інструменту MultiQC по набору даних TDP-43

Рід зразка	% Дуплікатів	% GC	Довжина	% Неуспішно	Млн рідів
Прямий	64.1%	51%	150 bp	20%	69.4
Зворотній	64.4%	51%	150 bp	20%	68.9

Було отримано високий відсоток дуплікатів – 64,1% для прямих та 64,4% для зворотніх рідів. Дуплікати в результатах секвенування - це копії фрагментів РНК, які були зчитані ідентичними або майже ідентичними послідовностями нуклеотидів. Дуплікати можуть виникати з різних причин,

наприклад, при поганій якості вхідного матеріалу, недостатньої кількості різних фрагментів, що зчитуються, або помилок в процесі ампліфікації. Наявність дуплікатів у результатах секвенування може призвести до спотворення результатів аналізу, зменшення точності і достовірності отриманих даних. Тому під час аналізу результатів секвенування дуплікати зазвичай відфільтровуються або оброблюються спеціальними алгоритмами для їх виявлення та усунення. В даному випадку високий відсотковий вміст дуплікатів пояснюється високою інтенсивністю («глибиною») секвенування та великою сумарною кількістю рідів – більше 68 мільйонів, - з-поміж яких більше 20 мільйонів рідів є унікальними, що є оптимальним результатом (рис. 3.1).

Довжина кожного із рідів становить 150 пар основ, що відповідає поставленим в експерименті параметрам і свідчить про відсутність помилок при зчитуванні фрагментів секвенування. Окрім цього, низький відсоток (20%) невдалих рідів також свідчить про успішний перебіг секвенування, що дозволило продовження роботи із обраним набором даних без їхньої корекції та додаткової обробки.

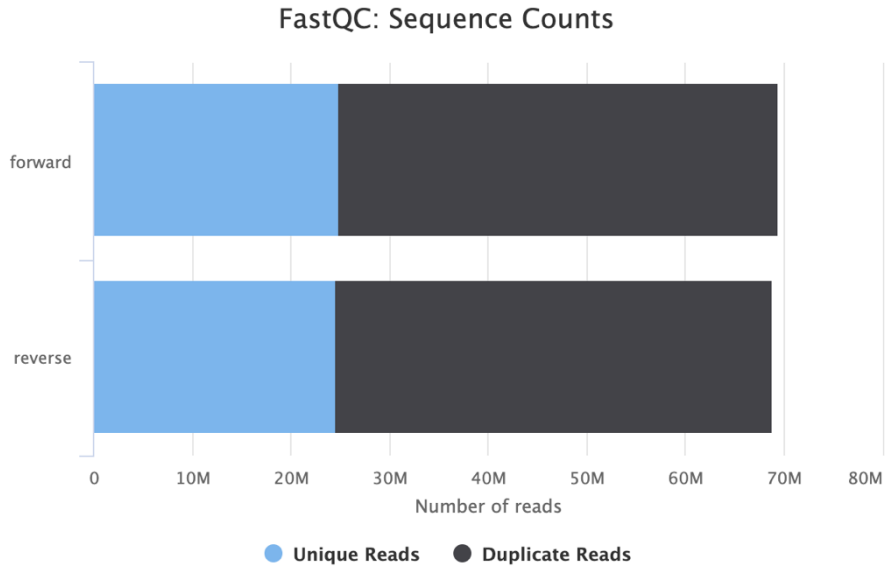

Рис. 3.1. Діаграма кількісної оцінки унікальних (**unique reads**) та дуплікованих (**duplicate reads**) рідів набору даних TDP-43 для прямих (forward) та зворотніх (reverse) рідів.

При усередненій оцінці якості для кожної пари основ в рідах перевірялось значення Phred score. Phred score - це числова оцінка якості зчитування нуклеотидної послідовності в результаті секвенування. Ця оцінка використовується для визначення ймовірності того, що нуклеотид, визначений в результаті секвенування, є правильно зчитаним (Zhang *et al.*, 2017). Phred score визначається за допомогою формули: phred score = -10 * log10(p), де p - ймовірність того, що нуклеотид визначений неправильно. Наприклад, phred score 20 відповідає ймовірності помилки 1 з 100 (або точності 99%), а phred score 30 - ймовірності помилки 1 з 1000 (або точності 99,9%). Phred score є важливим інструментом для оцінки якості результатів секвенування. Високі значення phred score свідчать про високу точність результатів, тоді як низькі значення можуть свідчити про помилки в процесі секвенування.

В даному випадку, значення phred score знаходилось в діапазоні вище 30 для всіх послідовностей, що означає 99,9% точність результатів секвенування (рис. 3.2).

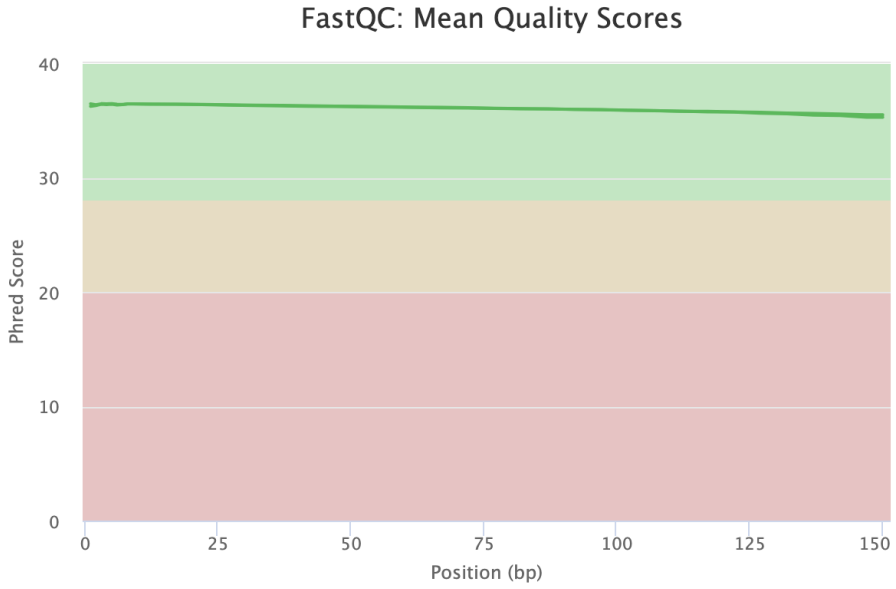

Рис. 3.2. Усереднені значення показника якості секвенування Phred score для кожного з нуклеотидів послідовностей.

Індикатори якості для усіх підмножин послідовностей також знаходяться в діапазоні оптимуму (рис. 3.3).

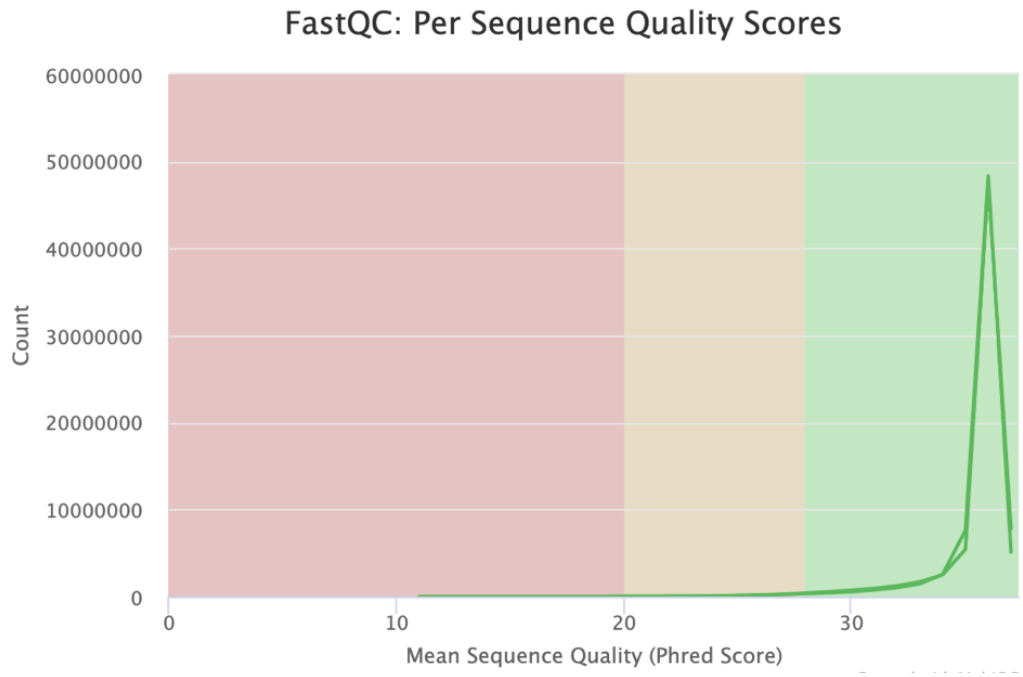

Рис. 3.3. Середні значення якості для підмножин послідовностей набору даних TDP-43.

Під час секвенування РНК, до бібліотеки даних зазвичай додаються адаптери - короткі послідовності ДНК або РНК, які використовуються для з'єднання з платформою секвенатора. Адаптери можуть бути додані до кінців фрагментів ДНК або РНК, що впливає на якість і точність результатів секвенування. У випадках, коли «забруднення» послідовностей адаптерами є високим, до даних застосовується процедура тріммінгу («відрізання») адаптерів. Тріммінг адаптерів полягає у видаленні цих послідовностей з рідів, отриманих під час секвенування, з метою забезпечення точності та якості даних (Dodt *et al.*, 2012). Недостатній тріммінг може призвести до помилкових позитивних результатів, що може впливати на інтерпретацію даних, особливо в випадку досліджень мутацій, генетичних варіантів і експресії генів. Одним із наслідків недостатнього тріммінгу адаптерів може бути поява збільшеної кількості помилок у послідовностях, що може спричинити проблеми при

аналізі даних за допомогою біоінформатичних інструментів. При обробці даних TDP-43 відсотковий вміст адаптерних послідовностей не перевищував 3%, що вказує на достатню якість послідовностей (рис. 3.4). Обидва ріди, прямий та зворотній, знаходяться в зонах оптимуму в контексті присутності адаптерних послідовностей. Як видно на діаграмі рисунку 3.4, вміст залишкових адаптерів зростає із довжиною рідів, що відповідає теоретичним очікуванням, оскільки приєднання адаптерів для секвенування відбувається саме на кінці послідовностей.

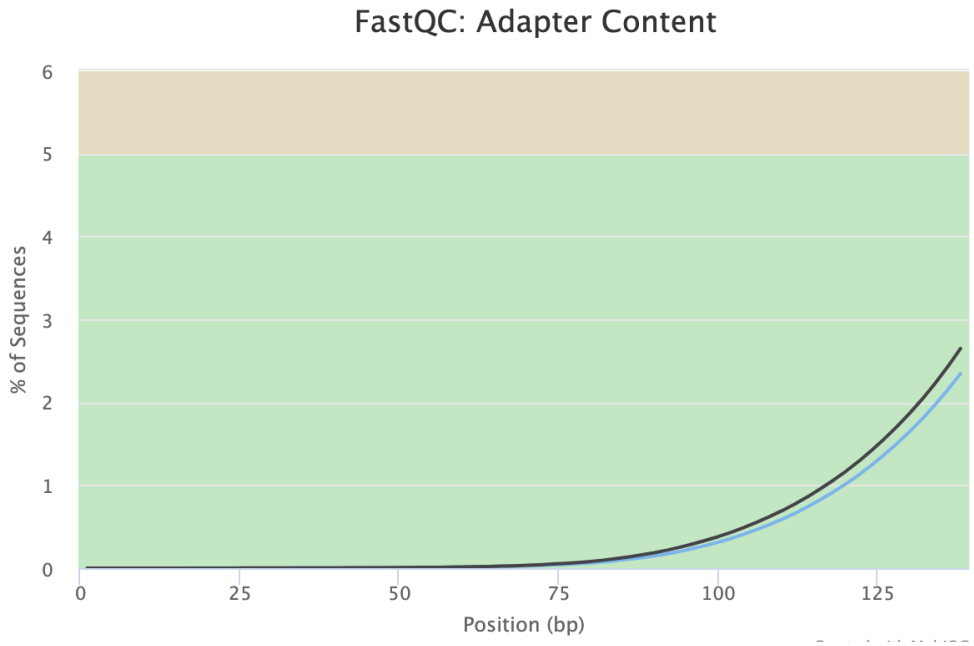

Рис. 3.4. Діаграма відсоткового вмісту адаптерних послідовностей для даних TDP-43.

Підсумовуючи описані вище результати, варто підкреслити, що отримані звіти для колекції даних можуть мати незначні відхилення при розгляді окремих зразків в індивідуальному порядку. Однак, зважаючи на високі показники якості та знаходження спостережуваних індикаторів в зонах оптимуму, результати перевірки якості для колекції послідовностей можна вважати оптимальними. Після застосування інструментів з обробки якості даних, набір зразків TDP-43 було завантажено до програмного алгоритму STAR з метою проведення геномного вирівнювання послідовностей.

3.1.2. Геномне вирівнювання послідовностей даних TDP-43 та кількісний підрахунок знайдених генів

Геномне вирівнювання послідовностей просеквенованого геному з референсним геномом є важливим етапом в аналізі даних секвенування. Референсний геном є стандартом, з яким, в даному випадку, порівнюються досліджувані набори даних. Референсний геном представляє собою високоякісну інформацію про послідовності генів та інших функціональних елементів (Formenti *et al.*, 2022). Основна мета геномного вирівнювання до референсного геному полягає в тому, щоб знайти розбіжності між відповідними регіонами, що порівнюються. Це може допомогти виявити гени, які були втрачені або набули нових функцій відносно референсного геному, а також - виявити мутації, що можуть бути пов'язані із певними патологічними проявами. Крім того, геномне вирівнювання може допомогти виявити зміни в кількості та розташуванні генів в досліджуваному наборі даних. Наприклад, відсутність або наявність дуплікатів генів, які можуть бути пов'язані зі специфічними фізіологічними характеристиками або захворюваннями. Також геномне вирівнювання може бути важливим інструментом для визначення відносної позиції генів та інших функціональних елементів у геномі, що може допомогти вивченню структури та регуляції геномів.

В даній роботі геномне вирівнювання послідовностей за допомогою інструменту STAR було необхідним кроком обробки даних, що забезпечував можливість подальшого аналізу диференційної експресії генів, оскільки завдяки вирівнюванню послідовностей була отримана інформація щодо конкретного розташування кожного із генів. В якості референсного геному використовувалась версія геному людини 39 (GRCh38.p13), яку було завантажено до серверу Galaxy із бази даних GENCODE у форматі GFF3.

Після застосування інструменту MultiQC до результуючих файлів вирівнювання у форматі LOG було встановлено, що відсоток вирівняних

послідовностей серед усіх наявних перевищував 92%, що вказує на успішне проведення процедури вирівнювання (таб. 3.2).

Таблиця 3.2

Звіт результатів геномного вирівнювання послідовностей даних TDP-43 із референсним геномом людини версії 39 (GRCh38.p13)

	% Aligned	M Aligned
WT-3	92.3%	64.1
S375G-1	92.6%	62.7
S375G-2	92.7%	66.2
S375G-3	92.5%	63.8
S375E-1	92.4%	60.8
S375E-2	92.9%	60.8
S375E-3	92.6%	59.4
WT-1	92.8%	66.3
WT-2	92.6%	59.7

З кількісної оцінки вирівняних послідовностей також варто відмітити високі показники, що знаходяться в діапазоні від 59 до 66 мільйонів вирівняних рідів. Однак, щоб коректно репрезентувати результати геномного вирівнювання послідовностей, варто також звертати увагу на співвідношення унікально вирівняних послідовностей, та таких, що за рахунок певних помилок мають виражену похибку у вирівнюванні. При вирівнюванні даних TDP-43 було встановлено, що більше 90% послідовностей є унікально вирівняними (рис. 3.5).

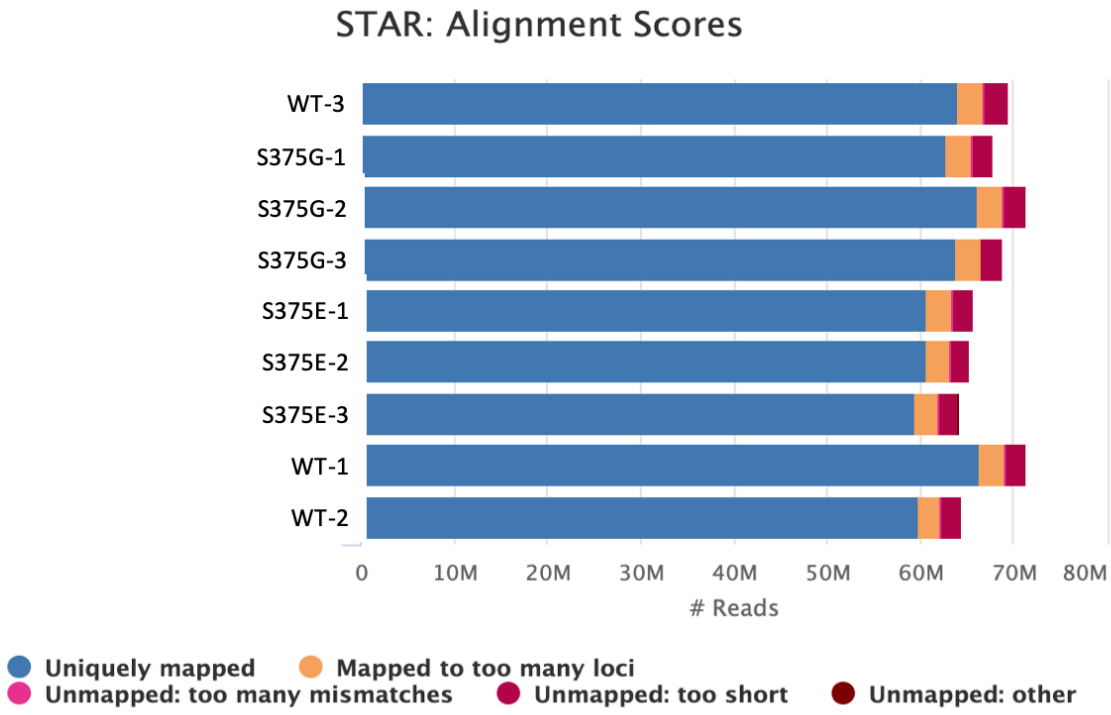

Рис. 3.5. Кількісні результати вирівняних послідовностей даних TDP-43. **Блакитний** – унікально вирівняні ріди; **оранжевий** – ріди вирівняні з великою кількістю локусів; **червоний** – невирівняні ріди, що мали занадто коротку довжину.

Таким чином, зважаючи на високий відсоток унікально вирівняних послідовностей, було зроблено висновок про успішне проведення процедури вирівнювання.

Для того, щоб кількісно визначити скільки саме рідів відповідають кожному із генів, було застосовано інструмент htseq-count, який передбачає вирівняні послідовності на попередньому кроці та референсний геном в якості вхідних даних, та формує таблицю по кожному із генів із відповідною кількістю співпадінь послідовностей зразків TDP-43 із послідовністю даного конкретного гена. Отримані файли з результатами підрахунку генів було імпортовано на локальний обчислювальний пристрій та завантажено до середовища RStudio з метою проведення подальшого дослідження зразків.

3.2. Обробка даних C9 HNRNPH1

Як було зазначено вище в підрозділі 2.3.2, дані просеквенованих зразків пацієнтів з діагнозом бічний аміотрофічний склероз з гексануклеотидними повторами в гені C9ORF72 (C9) та високим вмістом нерозчинних агрегатів білка HNRNPH1 попередньо вже були проаналізовані та оброблені до кроку геномного вирівнювання послідовностей в бакалаврській роботі за програмою «Біоінформатика» при Франкфуртському університеті ім. Й.В. Гете.

Для даної магістерської роботи було взято отримані файли з вирівняними послідовностями, після чого до них, аналогічно із даними TDP-43, було застосовано інструмент htseq-count з метою отримання підрахунку рідів для кожного із анотованих генів. Файли, що містять результати підрахунку генів, було імпортовано на локальний комп'ютер і завантажено до середовища RStudio для подальшого дослідження зразків.

3.3. TDP-43 аналіз диференційної експресії генів

Аналіз диференційної експресії генів передбачає дослідження зразків різного походження, наприклад, – контроль та мутаційна форма – на предмет особливостей вираженості певних генів. Відмінності в регуляції генів у мутаційних або патологічних зразках дозволяють зрозуміти роль гену або групи генів, які по-різному себе проявляють в здорових фізіологічних та мутантних / патологічних випадках.

Для того, щоб зрозуміти роль мутаційних форм S375E та S375G TDP-43, їхню експресію генів було порівняно із контрольними зразками. Після застосування алгоритму DESeq2 (Differential Expression analysis for Sequence Count data version 2) для обчислення параметрів диференційної експресії генів, було використано логарифічну трансформацію отриманих даних. Rlog (regularized logarithm) - це трансформація даних, яку використовують в аналізі

даних секвенування для виправлення дисперсії та нормалізації даних. Rlog трансформація даних застосовується для зменшення впливу великих значень на низькі, що підвищує точність аналізу. Вона є ефективнішою, ніж звичайний логарифм, оскільки регулює дисперсію високих та низьких значень, забезпечуючи більш надійну оцінку генетичної варіації. Rlog трансформація даних є одним з методів нормалізації та процесингу даних, які використовуються в аналізі диференційної експресії генів за допомогою DESeq2. У вигляді матриці дистанцій, rlog трансформація даних представляється як матриця відстаней між зразками, в якій значення відповідають відстаням між зразками у просторі ознак після застосування rlog трансформації даних. Ця матриця може використовуватися для кластеризації та візуалізації схожості між зразками у вигляді графіків, теплових карт або інших методів візуалізації даних.

Зразки S375E (TDP-43-S375E), S375G (TDP-43-S375G) та контрольні дані (TDP-43-CT) було порівняно між собою за параметром матриці дистанцій на основі rlog трансформації (рис. 3.6).

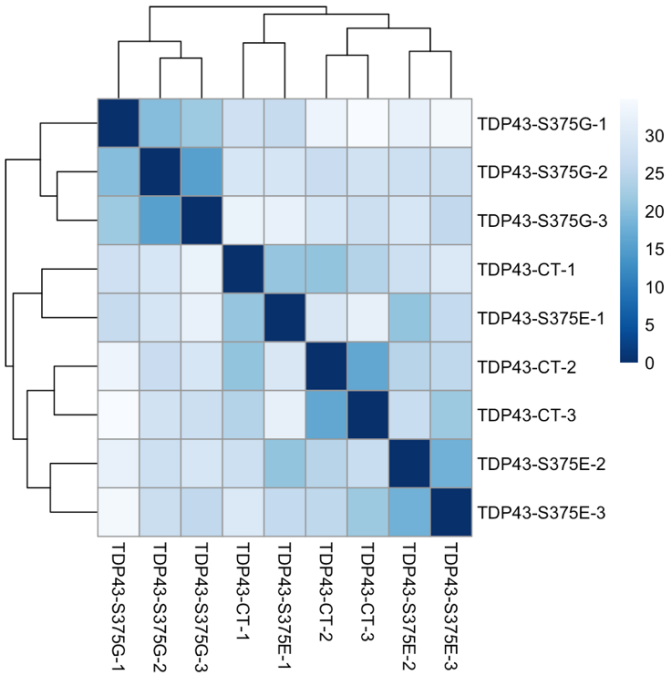

Рис. 3.6. Теплова карта порівняння зразків мутаційних форм TDP-43 S375E (TDP-43-S375E), S375G (TDP-43-S375G) та контрольних даних (TDP-

43-CT). Темно-сині області, які відповідають значенню 0 означають повну ідентичність порівнюваних зразків, в той час як світлі регіони теплової карти (зростаючі значення) вказують на високий ступінь різниці досліджуваних зразків.

Як видно на рисунку 3.6, досліджувані зразки кластеризуються у групи, що відповідають фізіологічним особливостям зразків. Так, найбільшу структурно-функціональну схожість мають три реплікати мутантних форм S375G, що вказує на високий рівень консервативності даної мутантної форми на рівні регуляції генів. Більш дисперсійне відхилення було зафіксоване для зразків S375E, а саме – для реплікатної форми під номером 1, що має певну відносну спорідненість із контрольним зразком №1. Така схожість іноді може пояснюватись певними варіативними відхиленнями при створенні ліній реплікатів, що можуть бути статистично незначущими.

В результаті аналізу за принципом основних компонент за 500 найбільш регульованими генами було досліджено кластеризацію зразків (рис. 3.7).

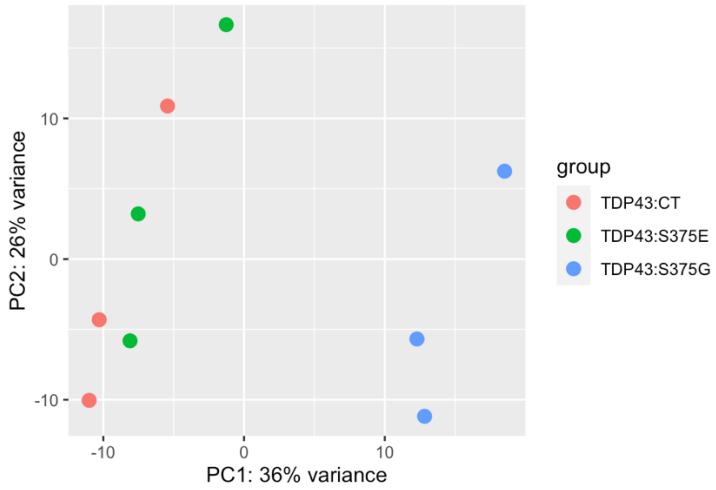

Рис. 3.7. Кластеризація зразків TDP-43 за результатами компонентного аналізу за 50.000 генами.

Результати аналізу за принципом компонент корелюють із отриманою візуалізацією на основі теплової карти. Найбільш наближено один до одного

розташовані зразки мутаційної форми S375G (рис. 3.7, **блакитний колір**), в той час як **контрольні зразки** та **мутантна форма S375E** характеризуються більшою подібністю між собою. Також помітною є різниця між двома мутантними формами, що вказує на різні механізми регуляції експресії генів при цих патологічних відхиленнях.

3.3.1. Особливості регуляції генів зразків S375E та S375G TDP-43

Для того, щоб зрозуміти особливості диференційної експресії генів кожну із мутаційних форм, TDP-43 S375E та S375G, було спочатку розглянуто окремо та порівняно із контрольним зразком, а потім - досліджено особливості регуляції генів у порівнянні між двома мутаційними формами.

Теплові карти двох мутаційних форм окремо зберігають тенденцію по відношенню до контрольної групи, яку було описано раніше, а саме – більш виражена консервативність та відмінність групи S375G із контролем (рис. 3.8 Б), та менш виражена відмінність групи S375E із контролем (рис. 3.8 А).

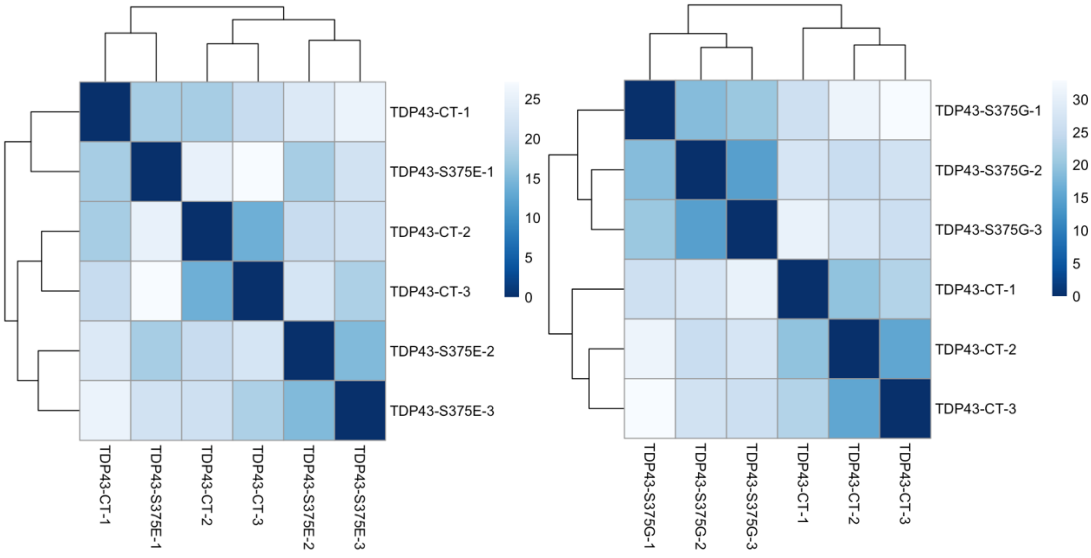

Рис. 3.8. Теплові карти логарифмічної трансформації зразків за принципом їхньої схожості / відмінності. (А) Порівняльна карта мутантної форми S375E та контрольних зразків. (Б) Порівняльна карта мутантної форми S375G та контрольних зразків.

Результати аналізу за принципом основних компонент дають більш зрозумілу закономірну кластеризацію мутантних форм у порівнянні із контрольними зразками (рис. 3.9).

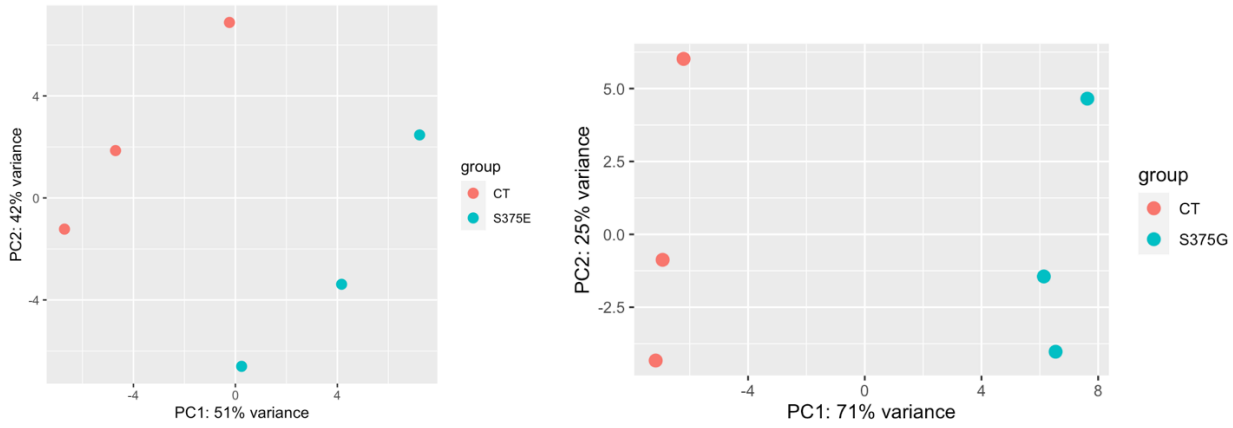

Рис. 3.9. Результати аналізу зразків за принципом основних компонент. (А) Кластеризація зразків мутантної форми S375E (**блакитний**) та контрольних зразків (CT, **червоний**). (Б) Кластеризація зразків мутантної форми S375G (**блакитний**) та контрольних зразків (CT, **червоний**).

Взаємовіддалене розташування контрольних і мутантних зразків відносно однин одного вказує на принципові відмінності в механізмах регуляції експресії генів в контрольних (фізіологічних) формах TDP-43 та при його мутаційних формах.

З-поміж загальної кількості знайдених генів було відокремлено лише ті, які статистично відрізнялись своєю регуляцією між мутантними та контрольними групами (рис. 3.10).

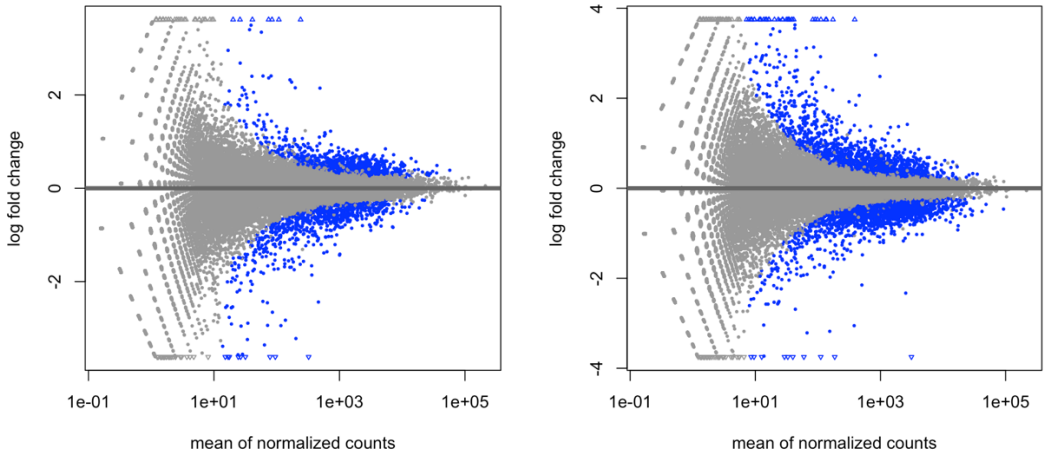

Рис. 3.10. Диференційна регуляція генів мутантних форм S375E (А) та S375G (Б) відносно контрольної групи. Значення log2Fold change в додатній області відповідають підвищеній експресії певного гена в мутантній формі у порівнянні із контролем, в той час як від'ємні значення log2Fold change вказують на пригнічену регуляцію певного гена в мутантній формі (або ж, якщо сформулювати це по-іншому, то підвищену регуляцію певного гену у контрольному зразку у порівнянні із мутантною формою). P-значення ≤ 0,05 (**синій**), p-значення > 0,05 (**сірий**).

Таким чином, кожен з елементів (крапок) графіків розподілу на рис. 3.10 відповідає певному гену. Сірий колір маркування відповідає статистично незначущій зміні експресії даного гена, в той час як синє маркування вказує на статистично обґрунтовану регуляцію розглянутого гену.

Щоб дати більш детальну кількісну оцінку диференційній регуляції генів між мутантними формами TDP-43 та контролем, для всієї множини генів було створено фільтр, що мав параметри |log2Fold change| > 1 та p-значення < 0,05. Таким чином, з усієї множини генів було підраховано лише ті, що мають статистично обґрунтовану регуляцію та змінений вектор експресії (рис. 3.11).

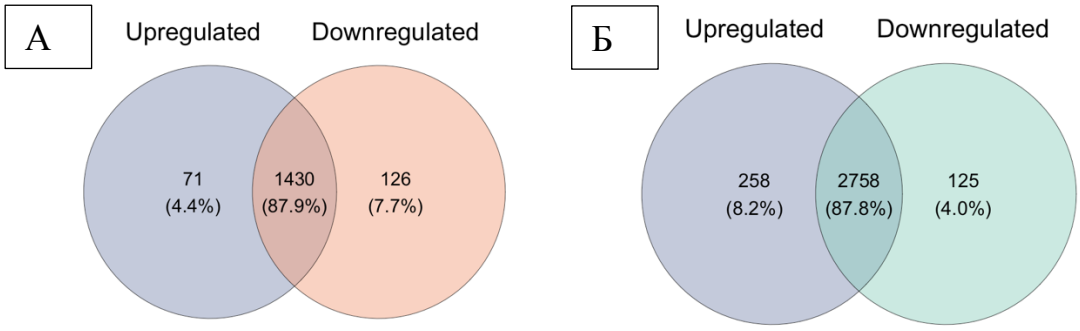

Рис. 3.11. Кількісна візуалізація генів з підвищеною (Upregulated) та зниженою експресією (Downregulated) у порівнянні мутантних форм TDP-43 S374E (А) та S375G (Б) із контрольними зразками.

З даних діаграм можна зробити висновок, що кількісний розподіл диференційної експресії генів для двох форм S375E та S375G є дещо різним. Так, наприклад обидві форми майже не відрізняються між собою за кількістю генів зі зниженою регуляцією (126 та 125 для S375E та S375G відповідно), але суттєво відхиляються від цієї тенденції при розгляді генів із підвищеною експресією – 71 та 258 генів для форм S375E та S375G відповідно. Також, якщо подивитись на переріз діаграм, то видно, що на проміжку значень, де Log2Fold change > -1 та Log2Fold change < 1, більша кількість статистично значущих генів була знайдена для групи S375G, що вказує на наявність неоднорідності та різниці механізмів регуляції експресії генів у двох мутантних формах TDP-43 S375E та S375G.

3.4. Порівняння регуляції генів при мутаційних формах TDP-43 та патології в гені C9 із високим вмістом нерозчинного білку HNRNPH1

Нейродегенеративні захворювання пов’язані із великою кількістю механізмів та факторів патогенезу. Часто при дослідженнях таких захворювань як хвороба Альцгеймера, бічний аміотрофічний склероз, лобно-скронева деменція та ін. прослідковується комплексна регуляція

патологічного розвитку, що перетікає під впливом одразу декількох чинників та характеризується багатофакторними порушеннями. Зокрема, часто розглядається взаємодія таких елементів патогенезу як зміни функціонування TDP-43 та наявність скупчень білкових агрегатів у вигляді включень, наприклад, HNRNPH1. Щоб дослідити можливі закономірності взаємодії даних факторів патогенезу на рівні регуляції експресії генів, зразки клітинних ліній мутантних форм TDP-43 було порівняно зі зразками пацієнтів із діагнозом бічного аміотрофічного склерозу, в яких спостерігалось накопичення нерозчинних агрегатів білку HNRNPH1.

При порівнянні зразків TDP-43 S365E та S375G зі зразками C9-High на основі матриці дистанцій логарифмічної трансформації було встановлено суттєву різницю між ними (рис. 3.12).

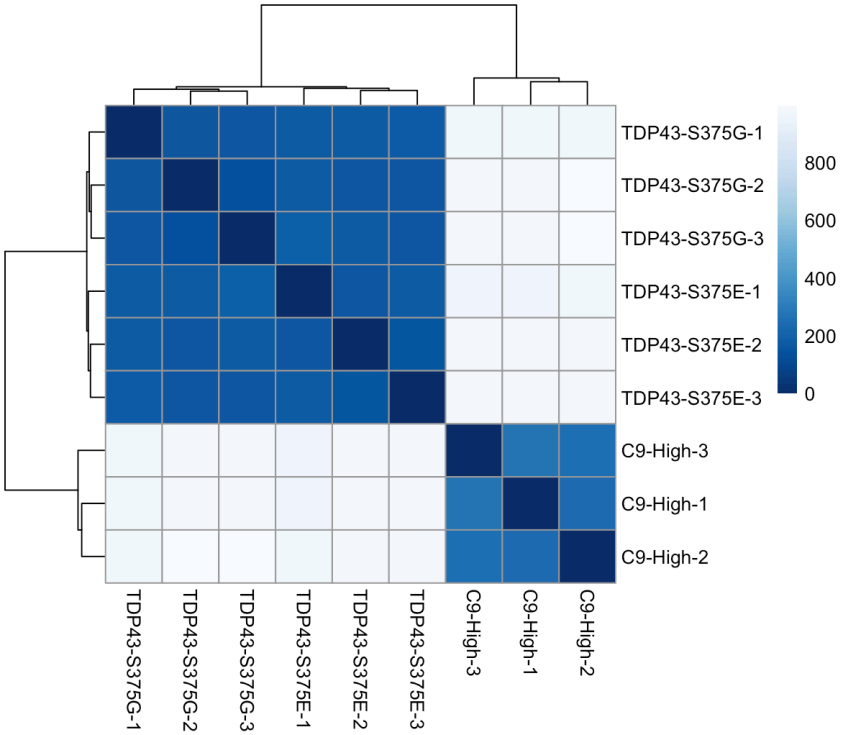

Рис. 3.12. Теплова карта порівняння зразків TDP-43 та C9-high.

Як видно на рисунку 3.12, при порівнянні зразків TDP-43 зі зразками C9-high прослідковується значна різниця між ними, що відображається в тепловій карті на основі матриці дистанцій. Така різниця є закономірним наслідком

різного походження зразків: експерименти TDP-43 мають походження із клітинної лінії, в той час як C9-high було отримано зі зразків мозку людей із діагнозом бічний аміотрофічний склероз. Різне походження зразків пов'язано передусім зі значною різницею в регуляції молекулярно-фізіологічних процесів, що також результується у прояві гетерогенності експериментів TDP-43 та C9-high (рис. 3.13).

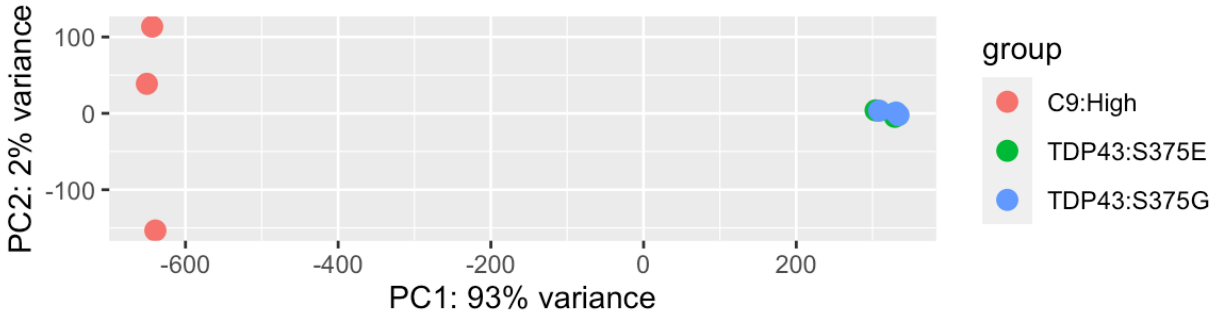

Рис. 3.13. Діаграма кластеризації зразків TDP-43 та C9-high за принципом компонентного аналізу.

Так само як і на тепловій карті, аналіз за принципом основних компонент відображає визначену кластеризацію зразків на дві групи: TDP-43 та C9-high. Помітно, що зразки групи TDP-43 мають більш щільне розташування у порівнянні із C9-high, що може бути пояснено складністю та неоднорідністю зразків мозку пацієнтів, комплексність яких може ускладнюватись індивідуальними фізіологічними особливостями.

Для того, щоб дослідити особливості схожих тенденцій регуляції експресії генів між експериментальними групами TDP-43 та C9-high, було встановлено відповідні параметри фільтрів результатів, а саме: гени, експресія яких вважалась підвищеною при патологічному стані, мали зміну log2Fold ≥ 0,05, в той час в якості генів зі зниженим рівнем експресії розглядались такі, що мали показник log2Fold ≤ -0,05. Значення p було встановлено менше 0,05. Таким чином, після застосування фільтру результатів було отримано карти

порівняння диференційної експресії генів для груп C9-high, S375E (TDP-43) та S375G (TDP-43) (рис. 3.14).

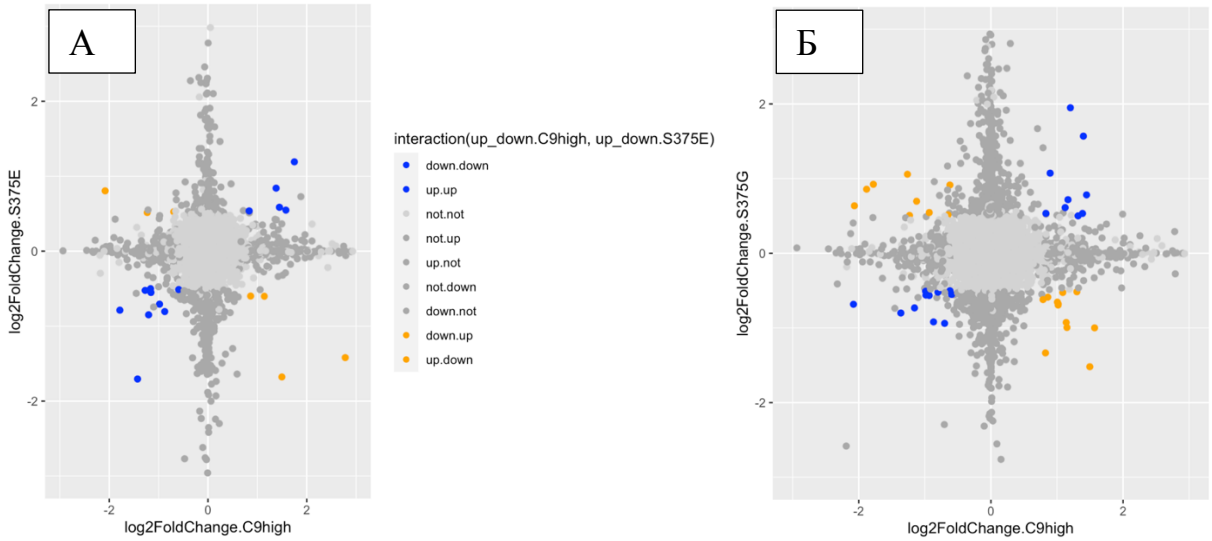

Рис. 3.14. Диференційна експресія генів при дослідженні експериментальних зразків патології C9-high та TDP-43 S375E (А) і TDP-43 S375G (Б). **Синім кольором** позначені гени з односпрямованою (підвищеною або зниженою) експресією в обох групах C9-high та TDP-43; **оранжевим кольором** позначені гени із різноспрямованою регуляцією генів між двома зразками. **Сірим кольором** позначені гени із неоднорідною або статистично незначущою регуляцією експресії.

Як видно на рисунку 3.14, гени, що розташовані в секторах I та III координатних площин (**синій колір**), мають однакову тенденцію регуляції для експериментів C9-high та TDP-43. Гени, що розташовані в секторі I, мають позитивні значення зміни log2Fold для обох груп, що означає їхню підвищену експресію в обох патологічних станах. Аналогічно цьому, гени в секторі III мають від'ємні значення зміни log2Fold, що вказує на знижений рівень їхньої експресі в C9-high та TDP-43.

З-поміж генів, які мають підвищену експресію в трьох зразках (C9-high, S375E, S375G) було знайдено 2 результати, притаманні кожній із груп.

Аналогічно цьому, серед генів зі зниженою експресією було визначено 3 представники (рис. 3.13).

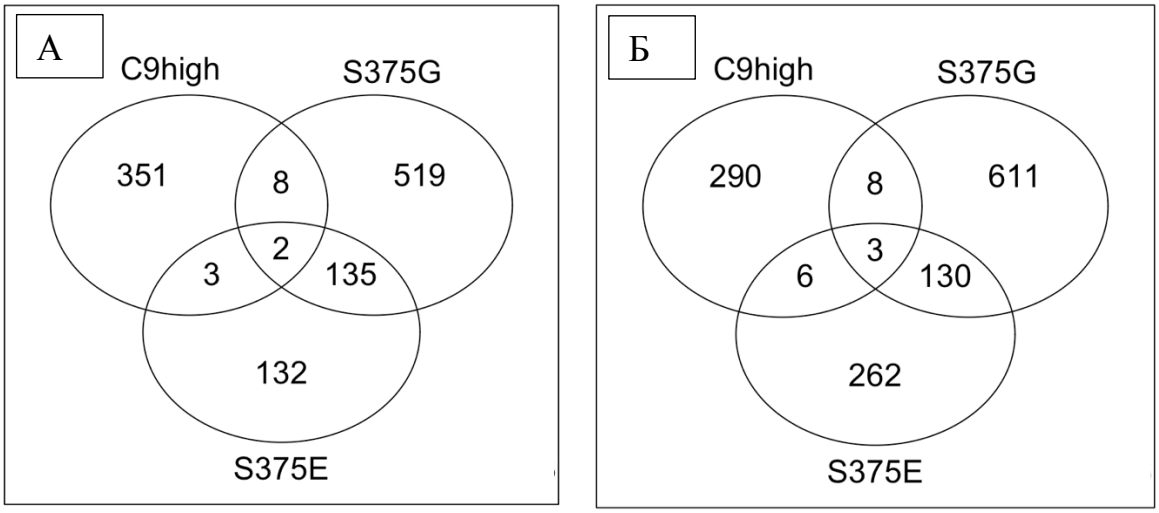

Рис. 3.14. Діаграма кількості генів із підвищеною (А) та зниженою експресією генів (Б).

Генами, які продемонстрували підвищену експресію в групах C9-high, S375E та S375G є NAALAD2 (індекс бази даних ENSEMBL: ENSG00000077616) та CCAR2 (індекс бази даних ENSEMBL: ENSG00000158941). Спільними генами зі зниженою експресією є: SLC39A9 (ENSG00000029364), PPM1A (ENSG00000100614) та FBXO34 (ENSG00000178974). Щоб детально зрозуміти та пояснити потенційний взаємозв’язок отриманих результатів із патогенезом нейродегенеративних захворювань, було детально розглянуто функцію деяких із визначених генів.

3.4.1. Особливості гену CCAR2 та його роль при патогенезі нейродегенеративних захворювань

Ген CCAR2 (cell cycle and apoptosis regulator 2) є білок-кодуючим геном, що розташований у 8-ій хромосомі. Даний ген забезпечує активність

зв'язування комплексу РНК-полімерази II та інгібіторну активність цього ферменту. Задіяний у кількох клітинних процесах, включаючи апоптоз, регуляцію клітинного циклу, метаболізму клітинних білків, регуляцію передачі сигналів та транскрипції (López-Saavedra *et al.*, 2016). В умовах оксидативного стресу цей ген підтримує цілісність мітохондрій, індукуючи апоптоз. Посилення регуляції цього гена свідчить про те, що популяція нейронів піддається значному окислювальному стресу, який, у свою чергу, індукує апоптотичні сигнали (Dharshini, Taguchi and Gromiha, 2019).

Захворювання, пов'язані з CCAR2, включають енцефалопатію, важку неонатальну патологію, спричинену мутаціями MECP2, та рак молочної залози. Серед пов'язаних з CCAR2 регуляторних шляхів виділяють клітинні реакції на подразники та на тепловий стрес (Kim, Cheon and Kim, 2017). В одному із досліджень із застосуванням методу повноекзомного секвенування ген CCAR2 було визначено одним із 14 нових біомаркерів нейродегенеративних захворювань (Cardoso *et al.*, 2019). Окрім цього, також повідомляється про участь гену CCAR2 у якості зворотнього регулятора рецепторного каскаду реакцій, що активується компонентом тирозин-кіназного рецептору трансмембранних глікопротеїнів EPHB2. Перебіг даного каскаду реакцій, в якому CCAR2 через активацію BAG2 спричиняє убіквітинування AIG1, є притаманним патологічним сценарієм при пізній стадії хвороби Альцгеймера (Yeh, Chung and Chen, 2021). CCAR2 активує SIRT1, а SIRT1 пригнічує CCAR2. Ці гени посилено регулюються при хворобі Альцгеймера і беруть участь в апоптозі та реакції на окисний стрес (Dharshini, Taguchi and Gromiha, 2019). Також, ген CCAR2 розглядається як модулятор при патогенезі хвороби Паркінсона (Vozdek, Pramstaller and Hicks, 2022).

Результати вищезазначених досліджень підкреслюють важливість гену CCAR2 при патогенезі нейродегенеративних захворювань. В експериментальному досліджені патологій груп C9-high та TDP-43 спільна підвищена регуляція експресії гену CCAR2 дозволяє припустити, що даний ген має потенційні властивості до комплексної взаємодії із іншими

патологічними проявами нейродегенеративних захворювань, такими як наявність високого рівня нерозчинних агрегатів білку HNRNPH1 та мутації в гені TDP-43. Всі вищезазначені аргументи дозволяють розглядати CCAR2 в якості потенційного біомаркеру групи нейродегенеративних захворювань.

3.4.2. Особливості генів PPM1A і FBXO34 та їх зниженої експресії при патогенезі нейродегенеративних захворювань

PPM1A кодує Mg2+/Mn2+-залежну білкову фосфатазу A1. Білок, який кодується цим геном, належить до родини протеїнфосфатаз Ser/Thr PP2C. Відомо, що члени родини PP2C є негативними регуляторами шляхів клітинної відповіді на стрес. Фосфатаза PPM1A дефосфорилює та негативно регулює активність MAP-кіназ та кіназ MAP-кінази (Mazumdar *et al.*, 2019). PPM1A дефосфорилює білки на етапі посттрансляційної модифікації та бере участь у багатьох біологічних процесах, включаючи ріст клітин, імунну відповідь та процеси метаболізму, які тісно пов'язані з патогенезом хвороби Альцгеймера (Calsolaro and Edison, 2016).

PPM1A є регулятором стимулятору генів інтерферону та опосередковує вроджену імунну сигналізацію, яка бере активну участь у патогенезі захворювань центральної нервової системи (Li *et al.*, 2021). Також повідомлялося, що в дослідженні 2022 року було виявлено пригнічені рівні експресії та ферментативної активності PPM1A як у мозку пацієнтів з хворобою Альцгеймера, так і в мозку мишей 3×Tg-хвороба Альцгеймера із нокдауном PPM1A (Lv *et al.*, 2022). Повідомляється, що активація PPM1A є багатообіцяючою потенційною терапевтичною стратегією лікування хвороби Альцгеймера.

Ген FBXO34 (F-box protein 34), також як і PPM1A, є білок-кодуючим геном. Члени сімейства білків F-box, такі як FBXO34, характеризуються наявністю приблизно 40-амінокислотного мотива F-box (Rangwala *et al.*, 2021).

FBXO34 є білками з'єднання для убіквітинових лігаз Skp1-Cul1-FBP (SCF) типу E3, які керують убіквітинуванням багатьох білків (Randle and Laman, 2016). Убіквітиновані білки містяться в нейрофібрилярних клубках та олігомерних Aβ-бляшках, а мутація гена убіквітину-B+1 призводить до дегенерації нейронів, що асоціюється з погіршенням просторової пам'яті та хворобою Альцгеймера. Зокрема, убіквітин-протеасомна система і тау-фосфорилювання при хворобі Альцгеймера тісно пов'язані між собою (Gu *et al.*, 2022). Як і ген PPM1A, FBXO34 має виражену знижену експресію при патогенезі хвороби Альцгеймера. FBXO34 та інші чотири гени-кандидати були перевірені для побудови діагностичної моделі, яка може бути корисною для діагностики хвороби Альцгеймера в клінічних умовах. Також, повідомлялося про ген FBXO34 як нещодавно відкритий біомаркер нейродегенеративних розладів (Hiz Kurul *et al.*, 2022).

З отриманих результатів можна зробити висновок, що гени PPM1A та FBXO34, які проявляють виражену знижену експресію в експериментальних групах C9-high та TDP-43, мають важливе значення в якості потенційних модуляторів стратегій терапевтичного лікування нейродегенеративних захворювань, а також – в якості біомаркерів даних патологій.

ВИСНОВКИ

1. Відомо, що патології РНК-зв’язуючих білків і РНК є центральними аспетками багатьох поширених нейродегенеративних захворювань. Патологія може виникнути внаслідок зниження експресії білок-кодуючого гена, агрегації РНК-зв’язуючого білка, мутаціями амінокислотних послідовностей тощо. Зокрема, все частіше підтверджується взаємозв’язок порушень функціонування РНК-зв’язуючих білків зі зміненою експресією генів, що потенційно можуть розглядатись в якості біомаркерів нейродегенеративних захворювань. Зокрема, значна увага приділяється мутаційним формам ДНК/РНК-зв’язуючого білку TDP-43 та паттернам нерозчинних агрегатів РНК-зв’язуючого білку HNRNPH1 в тандемі із гексануклеотидними розширеннями в гені C9ORF72.
2. Було проведено дослідження набору даних TDP-43, що складався із двох мутантних форм S375E, S375G та контрольних зразків. Встановлено, що зразки секвенування мають високі показники якості та пройшли багатоетапну перевірку із застосуванням ряду інструментів біоінформатичного аналізу.
3. Зразки мутантних форм TDP-43 S375E та S375G значною мірою відрізняються від контрольних зразків TDP-43 за механізмами регуляції експресії генів. Окрім цього, така різниця спостерігалась і у порівнянні мутаційних форм між собою, що також свідчить про неоднорідність регуляторного розподілу патологічних форм білка.
4. При досліджені кореляційних особливостей зразків TDP-43 та C9-high було встановлено значну дистанцію між ними в контексті молекулярно-фізіологічних особливостей. В першу чергу такі результати зумовлені значною віддаленістю походження зразків, оскільки форми TDP-43 належать клітинній лінії, в той час як зразки C9-high являються просеквенованим матеріалом мозку людей, діагностованих бічним

аміотрофічним слерозом. Також, варто підкреслити, що, не зважаючи на закономірні прояви патології TDP-43 та агрегації HNRNPH1 за наявності гексануклеотидних повторів в гені C9, сукупність механізмів, з якими ці фактори нейродегенерації взаємодіють, є дуже обширним та не має обмежуватись односпрямованою кореляцією в механізмах регуляції експресії генів.

5. Серед загальної множини диференційно експресованих генів у зразках TDP-43 та C9-high було встановлено 2 гени із підвищеною та 3 гени зі зниженою експресією для обох наборів даних.
6. До генів із спільно підвищеною експресією в обох наборах даних відносить CCAR2, який бере участь в модуляції патогенезу хвороб Альцгеймера та Паркінсона; нещодавно почав розглядатись в якості нового біомаркера нейродегенеративних захворювань. Підвищений рівень експресії гена CCAR2 в зразках TDP-43 та C9-high підкреслює його важливість в регуляції патологічних процесів за наявності факторів-супутників нейродегенеративних процесів та підтверджує результати останніх досліджень.
7. Гени PPM1A та FBXO34, які проявляли знижений рівень експресії в патологічних зразках мутантних форм TDP-43 у порівнянні із контролем та зразках мозку людей із діагнозом бічного аміотрофічного склерозу, потенційно можуть розглядатись в якості важливих регуляторних біомаркерних елементів патогенезу нейродегенеративних захворювань, що зокрема також зумовлюється наявністю даних про їх тісний взаємозв’язок із патогенезом хвороби Альцгеймера.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

Afgan, E. *et al.* (2018) ‘The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update’, *Nucleic Acids Research*, 46(W1). doi: 10.1093/nar/gky379.

Arai, T. *et al.* (2006) ‘TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis’, *Biochemical and Biophysical Research Communications*, 351(3). doi: 10.1016/j.bbrc.2006.10.093.

Balendra, R. and Isaacs, A. M. (2018) ‘C9orf72-mediated ALS and FTD: multiple pathways to disease’, *Nature Reviews Neurology*. doi: 10.1038/s41582-018-0047-2.

Bampton, A. *et al.* (2020) ‘The role of hnRNPs in frontotemporal dementia and amyotrophic lateral sclerosis’, *Acta Neuropathologica*. doi: 10.1007/s00401-020-02203-0.

Barrett, T. *et al.* (2013) ‘NCBI GEO: Archive for functional genomics data sets - Update’, *Nucleic Acids Research*, 41(D1). doi: 10.1093/nar/gks1193.

Bennion Callister, J. and Pickering-Brown, S. M. (2014) ‘Pathogenesis/genetics of frontotemporal dementia and how it relates to ALS’, *Experimental Neurology*. doi: 10.1016/j.expneurol.2014.06.001.

Calsolaro, V. and Edison, P. (2016) ‘Neuroinflammation in Alzheimer’s disease: Current evidence and future directions’, *Alzheimer’s and Dementia*. doi: 10.1016/j.jalz.2016.02.010.

Cardoso, A. R. *et al.* (2019) ‘Essential genetic findings in neurodevelopmental disorders’, *Human genomics*. doi: 10.1186/s40246-019-0216-4.

Conlon, E. G. and Manley, J. L. (2017) ‘RNA-binding proteins in neurodegeneration: Mechanisms in aggregate’, *Genes and Development*. doi: 10.1101/gad.304055.117.

Coughlin, D. G. *et al.* (2019) ‘Most cases with Lewy pathology in a population-based cohort adhere to the Braak progression pattern but “failure to fit”

is highly dependent on staging system applied', *Parkinsonism and Related Disorders*, 64. doi: 10.1016/j.parkreldis.2019.03.023.

Cummings, J. and Miller, B. (2007) 'The human frontal lobes', *Functions and disor*.

Cunningham, F. *et al.* (2022) 'Ensembl 2022', *Nucleic Acids Research*, 50(D1). doi: 10.1093/nar/gkab1049.

Dharshini, S. A. P., Taguchi, Y. H. and Gromiha, M. M. (2019) 'Investigating the energy crisis in Alzheimer disease using transcriptome study', *Scientific Reports*, 9(1). doi: 10.1038/s41598-019-54782-y.

Dodt, M. *et al.* (2012) 'FLEXBAR-flexible barcode and adapter processing for next-generation sequencing platforms', *Biology*, 1(3). doi: 10.3390/biology1030895.

Dreyfuss, G., Kim, V. N. and Kataoka, N. (2002) 'Messenger-RNA-binding proteins and the messages they carry', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm760.

Dugger, B. N. and Dickson, D. W. (2017) 'Pathology of neurodegenerative diseases', *Cold Spring Harbor Perspectives in Biology*. doi: 10.1101/cshperspect.a028035.

Formenti, G. *et al.* (2022) 'The era of reference genomes in conservation genomics', *Trends in Ecology and Evolution*. doi: 10.1016/j.tree.2021.11.008.

Frankish, A. *et al.* (2021) 'GENCODE 2021', *Nucleic Acids Research*, 49(D1). doi: 10.1093/nar/gkaa1087.

Geuens, T., Bouhy, D. and Timmerman, V. (2016) 'The hnRNP family: insights into their role in health and disease', *Human Genetics*. doi: 10.1007/s00439-016-1683-5.

Gómez-Tortosa, E. *et al.* (2017) 'Familial primary lateral sclerosis or dementia associated with Arg573Gly TBK1 mutation', *Journal of Neurology, Neurosurgery and Psychiatry*. doi: 10.1136/jnnp-2016-315250.

Gu, X. *et al.* (2022) 'Hub Genes, Diagnostic Model, and Predicted Drugs Related to Iron Metabolism in Alzheimer's Disease', *Frontiers in Aging*

Neuroscience, 14. doi: 10.3389/fnagi.2022.949083.

Haeusler, A. R. *et al.* (2014) 'C9orf72 nucleotide repeat structures initiate molecular cascades of disease', *Nature*, 507(7491). doi: 10.1038/nature13124.

Hiz Kurul, S. *et al.* (2022) 'High diagnostic rate of trio exome sequencing in consanguineous families with neurogenetic diseases', *Brain*, 145(4). doi: 10.1093/brain/awab395.

Hofweber, M. and Dormann, D. (2019) 'Friend or foe-Post-translational modifications as regulators of phase separation and RNP granule dynamics', *Journal of Biological Chemistry*. doi: 10.1074/jbc.TM118.001189.

Jiang, R. *et al.* (2022) 'Statistics or biology: the zero-inflation controversy about scRNA-seq data', *Genome Biology*. doi: 10.1186/s13059-022-02601-5.

Jo, M. *et al.* (2020) 'The role of TDP-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies', *Experimental and Molecular Medicine*. doi: 10.1038/s12276-020-00513-7.

Kanaumi, T. *et al.* (2013) 'Non-neuronal cell responses differ between normal and Down syndrome developing brains', *International Journal of Developmental Neuroscience*, 31(8). doi: 10.1016/j.ijdevneu.2013.09.011.

Kansal, K. *et al.* (2015) 'Dementia phenotypes associated with TDP43 versus mixed TDP43/Alzheimer pathology', *International Psychogeriatrics*, 27. doi: http://dx.doi.org/10.1017/S1041610215002161.

Kim, W., Cheon, M. G. and Kim, J. E. (2017) 'Mitochondrial CCAR2/DBC1 is required for cell survival against rotenone-induced mitochondrial stress', *Biochemical and Biophysical Research Communications*, 485(4). doi: 10.1016/j.bbrc.2017.02.131.

Kim, W., Kim, D. Y. and Lee, K. H. (2021) 'Rna-binding proteins and the complex pathophysiology of als', *International Journal of Molecular Sciences*. doi: 10.3390/ijms22052598.

Lagier-Tourenne, C., Polymenidou, M. and Cleveland, D. W. (2010) 'TDP-43 and FUS/TLS: Emerging roles in RNA processing and neurodegeneration', *Human Molecular Genetics*, 19(R1). doi: 10.1093/hmg/ddq137.

Li, D. *et al.* (2022) 'An evaluation of RNA-seq differential analysis methods', *PLoS ONE*, 17(9 Septamber). doi: 10.1371/journal.pone.0264246.

Li, F. *et al.* (2021) 'cGAS-stimulator of interferon genes signaling in central nervous system disorders', *Aging and Disease*. doi: 10.14336/AD.2021.0304.

López-Saavedra, A. *et al.* (2016) 'A genome-wide screening uncovers the role of CCAR2 as an antagonist of DNA end resection', *Nature Communications*, 7. doi: 10.1038/ncomms12364.

Lv, J. *et al.* (2022) 'Miltefosine as a PPM1A activator improves AD-like pathology in mice by alleviating tauopathy via microglia/neurons crosstalk', *Brain, Behavior, and Immunity - Health*, 26. doi: 10.1016/j.bbih.2022.100546.

Marcos-Rabal, P. *et al.* (2021) 'Neurodegenerative Diseases: A Multidisciplinary Approach', *Current Pharmaceutical Design*, 27(30). doi: 10.2174/1381612827666210608152745.

Mazumdar, A. *et al.* (2019) 'The phosphatase PPM1A inhibits triple negative breast cancer growth by blocking cell cycle progression', *npj Breast Cancer*, 5(1). doi: 10.1038/s41523-019-0118-6.

Meneses, A. *et al.* (2021) 'TDP-43 Pathology in Alzheimer's Disease', *Molecular Neurodegeneration*. doi: 10.1186/s13024-021-00503-x.

Michael, W. M., Eder, P. S. and Dreyfuss, G. (1997) 'The K nuclear shuttling domain: A novel signal for nuclear import and nuclear export in the hnRNP K protein', *EMBO Journal*, 16(12). doi: 10.1093/emboj/16.12.3587.

Nussbacher, J. K. *et al.* (2019) 'Disruption of RNA Metabolism in Neurological Diseases and Emerging Therapeutic Interventions', *Neuron*. doi: 10.1016/j.neuron.2019.03.014.

Nwakuya, M. T. and Biu, E. O. (2019) 'AN ILLUSTRATION OF PRINCIPLE COMPONENT ANALYSIS AS A', *International Journal of Scientific Research and Innovative Technology*, 6(6).

Paron, F. *et al.* (2022) 'Unraveling the toxic effects mediated by the neurodegenerative disease–associated S375G mutation of TDP-43 and its S375E phosphomimetic variant', *Journal of Biological Chemistry*, 298(8). doi:

10.1016/j.jbc.2022.102252.

Prashad, S. and Gopal, P. P. (2021) 'RNA-binding proteins in neurological development and disease', *RNA Biology*. doi: 10.1080/15476286.2020.1809186.

Randle, S. J. and Laman, H. (2016) 'F-box protein interactions with the hallmark pathways in cancer', *Seminars in Cancer Biology*. doi: 10.1016/j.semcancer.2015.09.013.

Rangwala, S. H. *et al.* (2021) 'Accessing NCBI data using the NCBI sequence viewer and genome data viewer (GDV)', *Genome Research*, 31(1). doi: 10.1101/gr.266932.120.

Renton, A. E. *et al.* (2011) 'A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD', *Neuron*, 72(2). doi: 10.1016/j.neuron.2011.09.010.

Scaber, J. and Talbot, K. (2016) 'What is the role of TDP-43 in C9orf72-related amyotrophic lateral sclerosis and frontemporal dementia?', *Brain*. doi: 10.1093/brain/aww264.

Srinivasan, K. A., Virdee, S. K. and Mcarthur, A. G. (2020) 'Strandedness during cDNA synthesis, the stranded parameter in htseq-count and analysis of RNA-Seq data', *Briefings in Functional Genomics*, 19(5–6). doi: 10.1093/bfgp/elaa010.

Stephenson, J. *et al.* (2018) 'Inflammation in CNS neurodegenerative diseases', *Immunology*. doi: 10.1111/imm.12922.

Stupnikov, A. *et al.* (2021) 'Robustness of differential gene expression analysis of RNA-seq', *Computational and Structural Biotechnology Journal*, 19. doi: 10.1016/j.csbj.2021.05.040.

Vozdek, R., Pramstaller, P. P. and Hicks, A. A. (2022) 'Functional Screening of Parkinson's Disease Susceptibility Genes to Identify Novel Modulators of α-Synuclein Neurotoxicity in Caenorhabditis elegans', *Frontiers in Aging Neuroscience*, 14. doi: 10.3389/fnagi.2022.806000.

Wang, J. *et al.* (2018) 'A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins', *Cell*, 174(3). doi: 10.1016/j.cell.2018.06.006.

Yeh, S. J., Chung, M. H. and Chen, B. Sen (2021) 'Investigating pathogenetic mechanisms of alzheimer's disease by systems biology approaches for drug discovery', *International Journal of Molecular Sciences*, 22(20). doi: 10.3390/ijms222011280.

Zhang, S. *et al.* (2017) 'Estimating Phred scores of Illumina base calls by logistic regression and sparse modeling', *BMC Bioinformatics*, 18(1). doi: 10.1186/s12859-017-1743-4.