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

Завідувач кафедри молекулярної біотехнології та біоінформатики
доц. Нипорко Олексій Юрійович
Протокол №____ засідання кафедри
від “____” _____20___ р.

КЛІНАЛЬНІСТЬ ЧАСТОТ ВАРІАНТІВ ГЕНІВ РЕПАРАЦІЇ ДНК У ПОПУЛЯЦІЯХ *DROSOPHILA MELANOGASTER* ЄВРОПИ

Випускна кваліфікаційна робота магістра
студента спеціальності
091 Біологія
ОП «Біоінформатика та структурна біологія»
Терпила Іллі Сергійовича

Науковий керівник
к.б.н., асистент кафедри молекулярної біотехнології та
біоінформатики ІВТ
Войтешенко Іван Сергійович

Оцінка захисту роботи
_____

Робота виконана на базі кафедри загальної та медичної генетики ННЦ “Інститут біології та медицини” Київського національного університету імені Тараса Шевченка під керівництвом асистент, к.б.н. С.В. Серги

Київ – 2022 р.

АНОТАЦІЯ

Терпило І. С. Клінальність частот варіантів генів репарації ДНК у популяціях *Drosophila melanogaster* Європи - Випускна кваліфікаційна робота магістра за спеціальністю 091 Біологія ОП «Біоінформатика та структурна біологія».

У роботі було проведено аналіз клінальностей варіантів генів європейських популяціях *Drosophila melanogaster* шляхом побудови моделей лінійної регресії для визначення кореляції між частотами алелів генів та географічними широтою, довготою та висотою над рівнем моря. Встановлено 4 гени, які демонструють довготну клінальність: *Sox102F*, *mav*, *CG32572*, *Ten-a*. Дані результати можуть використовуватись для подальшого вивчення адаптивної мінливості *D. melanogaster* у європейському регіоні.

Ключові слова: клінальність, лінійна регресія, *D. melanogaster*.

ЗМІСТ

ВСТУП

Незважаючи на широке фенотипову та генетичну біорізноманітність, біологічні види часто проявляють ознаки адаптації до умов локального навколишнього середовища [39]. Це свідчить про сильну селекцію варіантів ДНК, що забезпечують місцеву адаптивність. Еволюційні біологи довго прагнули поєднати цю генетичну мінливість з фенотиповою мінливістю та пристосованістю серед природних популяцій. Одним з ефективних підходів виявився збір окремих особин уздовж географічних трансект - таких як широта, довгота або висота - що передбачувано варіюють у абіотичних (температура, вологість, УФ-опромінення) та біотичних (наприклад, видове різноманіття, рівень конкуренції) умовах.

Оцінка мінливості вздовж подібних трансект дозволяє ідентифікувати кліни, які часто характеризують як географічний градієнт кількісної фенотипової чи генотипової ознаки [1]. Обидва типи клін - ті що розташовані вздовж дискретних середовищ та безперервних - історично багато вивчались та оцінювались теоретично та емпірично.

Відбір зразків уздовж клінів надає ряд унікальних переваг та може потенційно пом’якшувати змішані демографічні фактори які важко контролюються, коли зразки збирають з поодиноких уривчастих локацій. Наприклад, потік генів має бути більш передбачуваним уздовж клінів, таким чином спрощуючи задачу відрізнити адаптивну від неадаптивної мінливості [2].

Кліни дозволяють моделювати залежності з точністю, яка недоступна при нерівномірному відборі зразків. Наприклад, клін вздовж берегової довготи зустрічається на різних континентах [35]. Подібні патерни диференціації, підтверджені на різній місцевості свідчать про паралельну адаптивність.

Таким чином, відкриття та вивчення нових клінів дозволяє дослідникам вивчати фундаментальні питання щодо природного добору та генетичного базису адаптивності [40].

З огляду на вищезазначене **метою** роботи було визначити чи наявні кореляції між частотами поліморфізмів в генах репарації ДНК дрозофіл та географічними координатами в популяціях *Drosophila melanogaster* Європи.

Для досягнення цієї мети були поставлені наступні **завдання**:

1. Систематизувати геномні дані, отримані в результаті проекту консорціуму DrosEU та узгодити їх із географічними показниками популяцій *D. melanogaster* Європи
2. Здійснити побудову лінійної регресії у кожному сайті вздовж всього геному *D. melanogaster* для моделювання залежності між частотою однонуклеотидних поліморфізмів та показниками географічної широти, довготи та висоти над рівнем моря.
3. Виконати перевірку результатів лінійної регресії для детекції поліморфізмів із статистично значущим показником кореляції та генів, до яких вони належать
4. Проаналізувати отримані гени на предмет їх функцій та взаємозв’язків у метаболічних системах *D. melanogaster*

РОЗДІЛ 1
ЯВИЩЕ ПРОСТОРОВОЇ МІНЛИВОСТІ НА ПРИКЛАДІ *DROSOPHILA MELANOGASTER*

1.1. *D. melanogaster* як модельний об'єкт в геноміці

Плодова муха *D. melanogaster* — це невелика двокрила комаха, що походить з півдня пустелі Сахара [3] та колонізувала всі континенти крім Антарктиди в якості синантропного виду [4]. Протягом останніх 15000–20000 років дрозофіла розширила свій ареал проживання до Європи та Азії і лише близько 200 років тому закріпилась в Австралії та Північній і Південній Америках [5]. Завдяки короткому життєвому циклу та простим умовам утримання *D. melanogaster* набула популярності як лабораторний модельний організм внаслідок експериментів Вільяма Касла, а пізніше Томаса Ханта Моргана на початку XX століття [6]. У той час, коли основні принципи спадковості ще залишались об'єктом запеклих дискусій, Морган використав *Drosophila* для експериментального підтвердження та доповнення фундаментальних законів менделівської генетики, що призвело до відкриття генів та їх розташування хромосомах. Ця робота склала основу нашого сучасного розуміння генетичних механізмів та була нагороджена Нобелівською премією [7]. Згодом, система *Drosophila* була вивчена більш детально та допомогла зробити вагомий внесок у розробку численних генетичних інструментів, таких як балансерні хромосоми, ген-специфічні нокаут-мутанти та інші трансгенні конструкти, включно із системою Gal4/UAS для генетичної експресії або системою CRISPR/Cas9 - технології редагування геному. Більше того, маючи компактний геном обсягом близько 180 м.п.н., *D. melanogaster* стала одним з перших

еукаріотів, чий геном був повністю відсеквенований, зібраний та анотований [8].

Окрім суттєвих результатів у галузі функціональної генетики, *Drosophila* як система також виявилася ефективною у популяційному генетичному аналізі. Феодосій Добжанський разом із співробітниками та студентами був одним із перших, хто систематично досліджував генетичну мінливість *Drosophila*, фокусуючись на хромосомних інверсіях. Його новаторська робота сформувала базове уявлення про еволюційні процеси, які становлять основу генетичної мінливості, а згодом заклала ґрунт для виникнення сучасної еволюційної біології [9].

Шляхом секвенування гену *Adh* у 11 лініях, зібраних у 5 природних популяціях, Хадсон отримав перші дані про поліморфізми послідовності ДНК плодової мухи, ідентифікувавши лише один несинонімічний поліморфізм із 43 однонуклеотидних поліморфізмів [10].

Ще на початку 1980-х років ряд методів, базованих на використанні рестрикційних ферментів, були застосовані для кількісної оцінки природної генетичної мінливості у кількох локусах *D. melanogaster* [11, 12]. Згодом був проведений перший аналіз відсеквенованих методом Сенгера ДНК фрагментів [13]. Ці дослідження дозволили сформувати загальну картину геномних закономірностей мінливості у послідовностях ДНК, виявивши значне різноманіття “мовчазних” нуклеотидних сайтів, несинонімічну різноманітність; більш рідкісні малі інсерції, делеції та мобільні генетичні елементи [14]. На основі нульової гіпотези нейтральної еволюції Хадсон et al. створив перший статистичний селективний тест, що базується на порівнянні поліморфізму та дивергенції: тест Хадсона-Крейтмана-Агуаде (HKA) [15] постулює, що всі гени повинні демонструвати однакове співвідношення внутрішньовидової мінливості до міжвидової дивергенції у нейтральних сайтах. Як розширення тесту HKA, Макдональд і Крейтман розробили новий тест для специфічної детекції

позитивної селекції на білкових послідовностях, вперше використаний для виявлення позитивної селекції в локусі *Adh* в *Drosophila*. З тих пір цей тест став загальноприйнятим тестом нейтральності [16] Теоретично, співвідношення несинонімічної до синонімічної дивергенції має дорівнювати співвідношенню несинонімічних до синонімічних поліморфізмів, якщо несинонімічні сайти нейтральні або шкідливі, але вище, якщо вони адаптивні. Достовірні докази адаптивної молекулярної еволюції згодом були отримані шляхом застосування тесту Макдональда-Крейтмана та методів, базованих на ньому [17, 18]. Нарешті, головне відкриття, зроблене в *D. melanogaster*, полягало в тому, що рівень мінливості нуклеотидів позитивно корелює з локальною швидкістю рекомбінації [19], що вказує на те, що відбір є одним з обмежувачів геномної різноманітності.

У підсумку, плодова муха *D. melanogaster* є оптимальним модельним організмом для вивчення нейтральної та адаптивної еволюції геному, оскільки вона демонструє швидку та широку адаптивність протягом коротких (менше 20 поколінь) проміжків часу у природних популяціях [20, 21], має потужні генетичні інструменти [22], добре анотований геном [23], а також дані про геномні поліморфізми.. Крім того, нещодавно було секвеновано геноми понад 25 видів споріднених до *Drosophila* [24]. Зокрема, порівняльний геномний аналіз 12 видів дозволив отримати нове фундаментальне уявлення еволюції геному [25] і призвів до проекту ModENCODE [26], що має на меті виявлення функціональних елементів у геномах *D. melanogaster* та *C. elegans*.

1.2. Феномен клінальності

Клін — це географічний градієнт фенотипової або генетичної ознаки в межах окремого виду. Мінливість подібних ознак може коливатись у масштабах від метрів до тисячі кілометрів. Особливо часто кліни зустрічаються серед видів, широко розповсюджених по різних континентах. Існують вагомі докази того, що природний добір відіграє центральну роль у формуванні клінів, частково тому, що просторова мінливість будь-якої ознаки у значній мірі окреслює зрушення в біотичному та абіотичному середовищі. Кліни відомі як «кошмар таксономістів та подарунок еволюціоністів», оскільки їхня еволюція свідчить про кілька спірних аспектів в екології та еволюції, таких як ступінь і характер природного добору, процес розповсюдження та потоку генів, історичну демографію та видоутворення [27].

Попри те, що “клін” як термін був сформульований Джуліаном Хакслі лише в 1938 році, дослідники протягом століть спостерігали плавну зміну ознак всередині виду. Як наслідок, наразі існує значна кількість задокументованих прикладів клінальної мінливості вражає, що включають кліни в морфології, фізіології, поведінці та генетичних локусах. Окремі морфологічні кліни настільки поширені, що їх можна назвати «законом» природи. Найдавнішим та найбільш суперечливим з подібних клінів є правило Бергмана, згідно якого розмір тіла особини збільшується по широтному градієнту. Даний патерн широко поширений серед різних видів ссавців, птахів та певних груп комах. Виглядає таким чином, що добір має діяти по широтному кліну розміру тіла, оскільки ця закономірність розвинулась та підтвердилась на кількох лініях організмів та на декількох континентах. Крім того, частка цих клінів розвивається впродовж відносно короткого проміжку часу. Наприклад, інвазійні популяції плодової мухи *Drosophila suboscura* в Північній Америці

розвинули широтний клін, подібний до того, що був зафіксований у Європі менш ніж за 20 років (Рис. 1.1.).

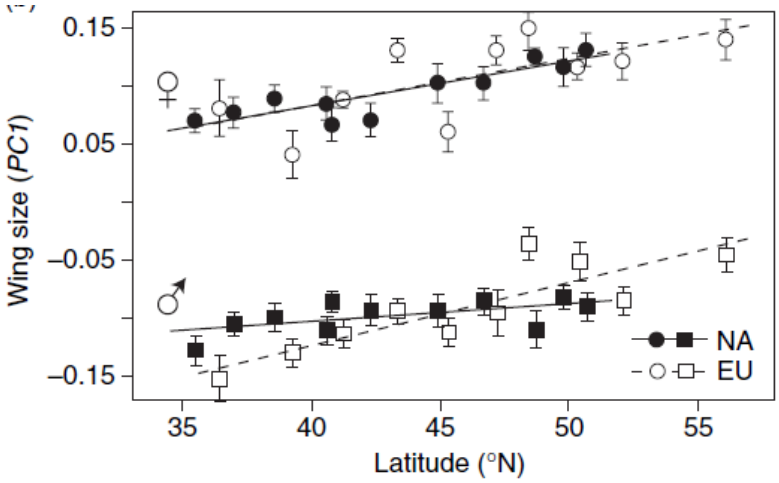

Рис. 1.1. Широтний клін розміру крил *D. obscura* в Європі та Північній Америці

Утім як і у випадку багатьох підтверджених клінальних патернів, селективні механізми, що лежать в основі правила Бергмана, залишаються не вивченими. Температура повітря різко знижується з широтою та є найбільш очевидним фактором впливу навколишнього середовища на розмір тіла, однак точний спосіб, через який імплементується залежність залишається невідомим. Серед інших поширених морфологічних клинів виділяють правило Аллена (пойкілотермні організми у холодному кліматі мають коротші виступаючі частини тіла) і правило Глогера (популяції у більш посушливих середовищах мають блідіший колір). Ці та інші кліни часто спірні. Наприклад, протягом 100 років після вторгнення в Північну Америку домашні горобці (*Passer domesticus*) розвинули клін за розміром тіла з півночі на південь (правило Бергмана) та клін за забарвлення зі сходу на захід за забарвленням (правило Глогера) [27].

Фізіологічні ознаки також проявляють клінальну мінливість. Наприклад, субпопуляції судинної рослини *Anthoxanthum odoratum* можуть

рости у ґрунті, контамінованому цинком. Ці рослини локалізовані на невеличкій ділянці менше 500 метрів вшир. Коли їх аналізують у лабораторії, виявляється, що фізіологічна толерантність рослин до цинку позитивно корелює з рівнем цинку в ґрунті. Ця закономірність вказує на селекцію цинк-толерантності у забрудненому грунті та чого вартує її підтримка у неконтамінованому грунті. Інші організми серед інших факторів навколишнього середовища проявляють клінальну мінливість фізіологічних реакцій на температуру, солоність та тривалість дня. Кліни ж поведінкових ознак часто зустрічаються у видів, де шлюбна поведінка для кожного з батьків змінюється залежно від простору [27].

Кліни в генетичних локусах (алозими, мікросателіти, білок-кодуючі локуси), зазвичай знаходяться під прямим добором або тісно пов’язані з локусами, що підлягають високому добору. Наприклад, лактатдегідрогеназа морської риби *Fundulus heteroclitus* змінюється по широті вздовж східного узбережжя США. Цей клін також був зафіксований серед людей у популяціях Європи та Близького Сходу [27].

Утім, деякі кліни можуть виникати в локусах, які не підлягають ані прямому, ні непрямому добору. Теоретично ці нейтральні кліни можуть бути наслідком вторинного контакту між популяціями (у випадку історичного розділення), які не перебувають у рівновазі. У рівновазі нейтральні клини повинні були б зникнути через триваючу інтрогресію алелів через потік генів [27].

1.3. Широтна клінальність *D. melanogaster*

Рід *Drosophila* представляє інтерес як система вивчення природного добору та демографії. Ця група складається з кількох широко поширених видів, які колонізували як Старий так і Новий Світи. *D. melanogaster* була і

продовжує детально вивчатись в контексті широтної мінливості. Ряд фенотипових ознак і генетичних локусів змінюються залежно від широти у *D. melanogaster* [42], також широтна мінливість *D. melanogaster* вивчалися в контексті геноміки на декількох континентах, таких як Північна Америка, Австралія, Європа, Азія та Африка. [43]. Однією з переваг використання *D. melanogaster* для вивчення широтної адаптації є те, що він лише недавно колонізував помірний клімат (10000–20000 років після виходу з Центральної Африки).

Ознаки, адаптовані до помірного клімату, такі як холодостійкість або резистентність до голоду повсюдно представлені у північних широтах стали результатом локальної адаптації до температурних кліматів [44]. Крім цього, спостерігається певний паралелізм закономірностей клінальності алельних частот уздовж Північно-Американських та Австралійських широтних клінів, що дає змогу висувати гіпотезу про конвергентну адаптацію до широти [43]. Широтні кліни *D. melanogaster* також піддаються певним змішаним демографічним ефектам. Популяції Північної Америки та Австралії видаються результатом змішування між Європейською та Африканською популяціями [45]. Незважаючи на те, що широтні кліни *D. melanogaster* доволі строгі, демографія ускладнює визначення алгоритму селекції.

Порівняльні дослідження можуть допомогти нам зрозуміти основні патерни широтної мінливості. Пара споріднених види *D. simulans* та *D. melanogaster* [46] представляють потужну систему для порівняльного генетичного систему. Ці види подібні у своєму ареалі проживання, екології та еволюційній історії [47]. Обидва види пройшли через подібну експансію із Африки, адаптацію до помірного клімату та розвиток коменсалізму з людьми [48]. На жаль, обмежений обсяг досліджень щодо клінальної мінливості *D. simulans* унеможливлює масштабне порівняльне дослідження широтної мінливості.

У той час як *D. simulans* демонструє клінальну мінливість ряду спільних ознак із *D. melanogaster*, таких як пігментація, розмір тіла, *D. simulans* також виглядає менш адаптованою до помірного клімату [49]. Наприклад, *D. simulans* має меншу фізіологічну стійкість до холоду та голодування [50]. Іншою ключовою клінальною ознакою *D. melanogaster* є репродуктивна діапауза, що є важливою для виживання під час холодних зим у північних широтах [51].

Широтна клінальність зустрічається в *D. melanogaster* не лише на рівні частот алелів, але і цілих хромосомних інверсій. Наприклад, великий (~8 м.п.н.), космополітичний інверсійний поліморфізм на правому плечі третьої хромосоми *In(3R)Payne,* також відомий як *In(3R)P*, демонструє клінальну мінливість за широтним градієнтом на кількох континентах, що особливо помітно вздовж східного узбережжя Австралії та Північної Америки [52]. Ця інверсія має африканське походження, виникла близько 130000–150000 років тому та довгий час вважалася важливою рушійною силою кліматичної адаптації *D. melanogaster* [53].

Потенційно постійний у ролі інверсії в кліматичний адаптації, інвертований каріотип In(3R)P демонструє середньо-високі частоти у південних широтах, тобто у субтропічному і тропічному кліматі та майже відсутній у північних - в помірному кліматі. Вздовж східного узбережжя Північної Америки, наприклад, інверсія має частоту близько 50% у штаті Флорида і падає майже до 0% у штаті Мен [52]. Згідно останніх даних висувається припущення, кліни в Австралії та Північній Америці адаптивно підтримуються шляхом просторового добору [28]. Незважаючи на те, що генетичні закономірності клінальних мінливостей можуть підлягати впливу змішування та вторинному контакту з предковими популяціями [54], північноамериканський клін в In(3R)P, гіпотетично

підтримується не нейтрально та зберігається незалежно від структури популяції та генетичного змішування [28].

1.3. Довготна клінальність *D. melanogaster*

На відміну від висотної чи широтної клінальності, градієнт за довготою в *D. melanogaster* зустрічається набагато рідше. Значущі довготні кліни були знайдені для деяких хромосомних інверсій, наприклад у ряді Африканських популяцій. В дослідженні Aulard et al. виявилось, що існує значуща кореляція чотирьох основних космополітичних інверсій по довготі, а саме: In(2L)t, In(2R)NS, In(3L)P and In(3R)P [55]. Після корекції по широті да висоті, дана закономірність підтвердилась для усіх інверсій окрім In(2R)NS. Найбільший ефект спостерігався для In(3L)P, у якій довгота впливала на 74% загальної мінливості. Для In(2L)t та In(3R)P довгота складала близько 36% географічної мінливості. У всіх випадках, частоти інверсій знижувались у напрямку з Заходу на Схід.

Відомостей про довготну клінальність алельних частот певних генів в *D. melanogaster* наразі немає. Утім, в інших видах *Drosophila* довготна клфнальність ознак підтверджена. Наприклад, в *D. americana* зафіксований довготний клін кольору тіла.

1.4. Висотна клінальність *D. melanogaster*

На відміну від широтних клінів, висотним клінам історично приділялось значно менше уваги дослідників, особливо у *Drosophila* [56]. Оскільки середня температура знижується як функція зростаючої широти та зростаючої висоти, висотні кліни часто сприймали як дзеркальні широтні кліні, принаймні якісно [56]. Справді, подібно до широтних клінів розміру тіла, висотні закономірності були зафіксовані у *D. buzzatii* [57], *D.*

robusta [58] і *D. takahashii* [59], а також розміру крил (як похідної розміру тіла) також у *D. melanogaster* [56]. Прогноз стосовно того, як висотні кліни відзеркалюють широтні не завжди підтверджувався. У той час як одні дослідження демонстрували позитивну залежність розміру тіла із висотою над рівнем моря, інші це спростовували [56]. Наприклад, результати дослідження диких мух не виявили відмінностей у розмірах тіла за висотним градієнтом [60]. Наразі, враховуючи те, що розмір тіла - це дуже пластична ознака, яка підлягає впливу багатьох чинників, важко зробити однозначні висновки, спираючись лише на дані польових спостережень. Таким чином, питання, чи є висотні кліни розміру тіла загальною закономірністю, що стосується більшості пойкілотермних організмів, або такі кліни є таксон-специфічними, досі залишається без достеменної відповіді.

Загалом, наше розуміння висотної мінливості генетичних ознак залишається обмеженим. Крім висотного зниження температури, швидкість якого становить приблизно 6-7°C/км у напрямку знизу догори, існує багато інших абіотичних факторів, що змінюються з висотою: зниження атмосферного тиску; підвищення рівня радіації та збільшення частки УФ-випромінювання [61]. Зниження тиску атмосферних газів, особливо кисню, може вплинути на швидкість метаболізму, росту та загальне виживання комах [60]. Крім того, знижена щільність повітря на великій висоті може суттєво погіршити польотні характеристики. Підвищений рівень радіації на великій висоті викликає певні відмінності між температурою атмосфери та певних ділянок земної поверхні, через що виникають особливі висотні мікроареали з вищою температурою поверхні порівняно з мікроареалами у північноширотних середовищах з аналогічною температурою повітря [60]. Крім того, на відміну від помірної біоти у високих широтах, високоширотна тропічна біота не стикається зі скороченням тривалості сезону. Додатковою важливою характеристикою

висотних градієнтів є те, що вони можуть спричиняти значні зміни навколишнього середовища на дуже коротких горизонтальних відстанях [59]. Це, у свою чергу, може призвести до високого рівня обміну генів між сусідніми районами, за рахунок чого висока генетична диференціація по висоті стає можливою за наявності умов сильного добору [62]. Загалом, усі ці різноманітні фактори можуть викликати відмінності в адаптації до висотних і широтних градієнтів, але ступінь таких відмінностей досі детально не встановлений.

1.5. Метод Pool-seq та його переваги

Технології секвенування нового покоління (англ. Next Generation Sequencing або NGS) наразі повсюдно використовуються для отримання даних про поліморфізми в модельних та немодельних організмах [63].

Спершу стратегія секвенування полягала в зборі та вивченні зразків окремих людей (The International HapMap Consortium 2005), однак на даний момент повногеномне секвенування груп особин або так званих “пулів” (Pool-seq) набуває все більшої популярності як метод в популяційній геноміці [64]. Оскільки Pool-seq працює з бібліотеками секвенування, що містять об’єднані групи зразків, та не потребує баркодного мічення кожного зразка, цей метод дає змогу отримати велику кількість повногеномних поліморфних даних за відносно невисоку ціну [64]. Утім не еквімолярні кількості ДНК з усіх зразків у пулі, а також стохастична мінливість ефективності ампліфікації окремих зразків у цьому пулі піднімають проблему точності отриманих оцінок алельних частот, особливо у випадку низької глибини секвенування та малого розміру складових пулів [63]. Тим не менш, було показано, що при однакових

зусиллях прикладених до секвенування, Pool-seq забезпечує подібні, якщо не точніші, оцінки частоти алелів, аніж індивідуальне секвенування.

Секвенування наступного покоління (NGS) здійснило революцію в дослідженнях з еволюційної генетики та мінливості за рахунок значного зниження технічних витрат та збільшення масштабу доступних даних на рівні цілого геному. Для повного розкриття потенціалу NGS також необхідно розробляти відповідні статистичні інструменти для нейтралізації недоліків цих даних, головним чином їх високий рівень помилок та їх дуже незбалансований характер. Останнім часом аналітичні методології, що використовуються для вивчення мінливості, прогресують, щоб максимально використати переваги NGS. Наприклад, кілька методів були зосереджені на отриманні припущення щодо еволюційного характеру зразків з використанням послідовності геному кількох або навіть однієї диплоїдної особини на популяцію [65].

Наразі використовується кілька ефективних стратегій для вивчення популяційної мінливості, таких як зменшення кількості ДНК для секвенування (англ. Reduced Representation Libraries; RRL) або секвенування ДНК асоційоване з сайтами рестрикції (Restriction-site Associated DNA Sequencing; RADSeq). RRL та RADSeq успішно застосовувались для вивчення еволюційних процесів у популяціях [66]. Поєднання кількох особин у пул виявилося дуже вигідним варіантом для NGS досліджень [67]. Іноді об’єднана вибірка є не варіантом, а необхідністю, як, наприклад, при вивченні поліплоїдних особин або дослідженні організмів, що живуть у щільних спільнотах (наприклад, коралів).

Іноді об’єднана вибірка (пул) виявляється не просто варіантом, а необхідністю, як, наприклад, при вивченні поліплоїдних особин або дослідженні організмів, що живуть у щільних спільнотах - наприклад, корали. Секвенуючи в пулах, ми втрачаємо можливість зв’язувати

послідовність з індивідом, якому вона належить. Утім для більшості досліджень з популяційної генетики, Pool-seq є вкрай ефективним методом зниження витрат на секвенування без зменшення розміру експериментальної вибірки.

1.6. Кореляція та лінійна регресія

Мета кореляції полягає в тому, щоб зрозуміти силу і характер взаємозв'язку між двома різними змінними. Сила такого відношення описує, наскільки тісно дві змінні асоційовані одна з одною, і вимірюється у діапазоні від 0 до 1, де 0 вказує на відсутність зв’язку, а 1 свідчить про наявність ідеальної лінійної залежності. У соціальних науках кореляція 0,1 зазвичай вважається слабкою, 0,3 вважається помірною, а >0,5 - сильною. Варто зазначити, що ці межі є умовними та відносними, тому вони варіюють у різних дисциплінах. Крім того, напрямок кореляції може бути як позитивним, так і негативним. Позитивна кореляція виникає, коли обидві змінні рухаються в одному напрямку - наприклад, зі збільшенням температури повітря на вулиці, зростає рахунок за використання кондиціонера.

Результати кореляційного аналізу позначаються літерою «r» і, яка має значення, що характеризує як силу кореляції (від 0 до 1), так і напрям (тобто + або -) зв’язку між двома змінними, що представляють інтерес. Високий коефіцієнт кореляції свідчить про існування лінійного зв’язку між двома змінними, низький же - навпаки, заперечує лінійну залежність. Також слабкий коефіцієнт кореляції може вказувати на криволінійну залежність двох змінних, через що потребуватиметься інший тип аналізу.

Слід зазначити, що самі по собі сильні кореляції не дорівнюють причинно-наслідковим зв’язкам.

Традиційний спосіб візуалізації кореляції між двома змінними - діаграма розсіювання, де кожна ось репрезентує одну змінну. При позитивній кореляції точки даних збільшуються від лівого нижнього кута у напрямку правого верхнього. Чим щільніше точки кластеризовані разом по діагоналі, тим сильніша кореляція. Коли ж не виникає лінійного патерна та точки просто довільно розміщені на графіку, це свідчить про відсутність кореляції.

Лінійні регресії можна розглядати розширенням кореляції Пірсона. У той час як кореляція визначає, чи існує зв’язок між двома змінними, регресія аналізує чи може незалежна змінна передбачити значення іншої залежної змінної. Для аналізу необхідно визначити, яка з двох змінна буде залежною (результат), а яка незалежною, також відомою як предиктор. На відміну від кореляційного аналізу, змінні у регресійному аналізі повинні бути класифіковані як результат та предиктор.

У простій лінійній регресії лише одна незалежна змінна використовується для передбачення однієї залежної змінної. Головна різниця між графіками регресії та кореляції полягає в тому, що графіки регресії також мають пряму лінію - лінію найліпшої відповідності. Лінія найліпшої відповідності демонструє основний напрямок та тренд розподілу даних. Стандартне рівняння простої лінійної регресії виглядає наступним чином:

$$Y_i = (b_0 + b_1 X_i) + e_i, \qquad (1.1)$$

де лінія найліпшої відповідності визначається кутом нахилу або коефіцієнтом регресії незалежної змінної (b_1) та точкою, в якій лінія перетинає вертикальну ось у (b_0). Змінна Y_i - залежна змінна для

передбачення, X_i - незалежна змінна, а e_i - похибка, що характеризується як відмінність між лінією найліпшої відповідності та справжніми даними

РОЗДІЛ 2
МАТЕРІАЛИ ТА МЕТОДИ ДОСЛІДЖЕНЬ

2.1. Матеріали та методи

В роботі був проаналізований файл у форматі VCF, що містив 48 Pool-seq зразків геному популяцій *D. melanogaster* із 32 локацій, розташованих по всій Європі [28]. Файл був люб'язно наданий для аналізу європейський консорціумом з популяційної геноміки дрозофіл DrosEU.

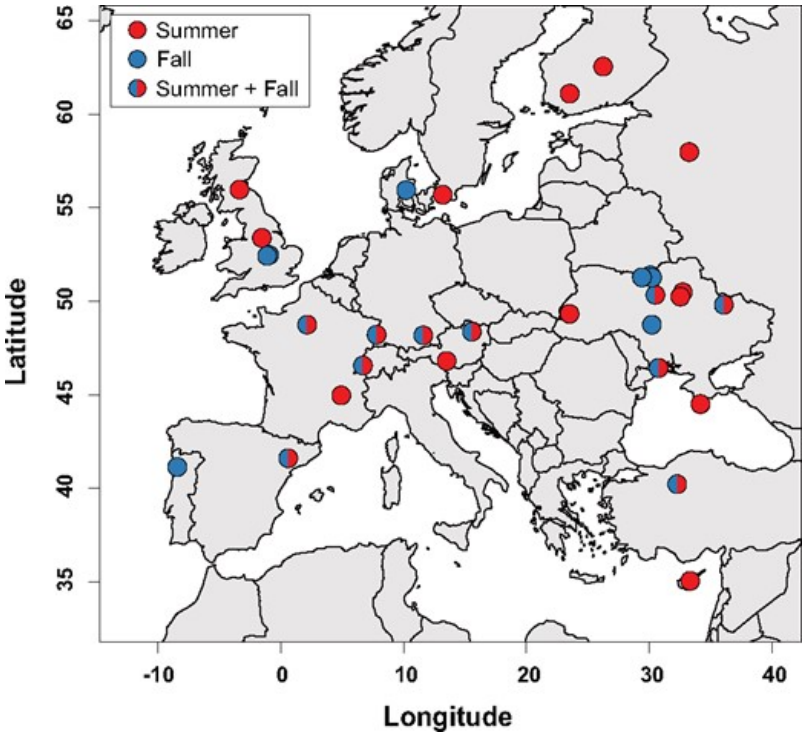

Рис. 2.1. Географічний розподіл зразків популяцій. Червоний колір позначає точки збору влітку, синій колір - точки збору восени. В 10 локаціях з 32 збір виконали в обидва сезони, вони позначені двома кольорами.

Також консорціумом DrosEU було надано дві допоміжні Excel-таблиці з інформацією про географічні показники кожної з популяцій (додаток 1).

2.2. Обладнання і програмне забезпечення

Аналіз був здебільшого виконаний на ноутбуці Acer Aspire 5 з наступними технічними характеристиками: процесор Intel Core-i7, 8 гб оперативної пам’яті, відеокарта NVIDIA GeForce MX150 з 2 гб відеопам’яті та SSD диск обсягом 256 гб. Для паралелізації обрахунків був залучений хмарний сервіс Google Colab [36].

Для поділу вихідного VCF архіву на хромосомні компоненти був використаний програмний застосунок SnpSift. Для аналізу VCF файлів використовували пакет vcfR [30] у середовищі RStudio 2021.09.2, мову програмування R 3.6, бібліотеку allel [38] у мові програмування Python, а також базу даних STRING 11.5 [37].

2.3. Підготовка робочої таблиці з географічними показниками

З вихідних даних була використана загальна таблиця у форматі Excel, що містить б усю необхідну інформацію для аналізу, а саме: географічні широта, довгота та висота над рівнем моря, сезон збору, кількість особин та хромосом у кожному пулі.

Дана таблиця була залучена до подальших програмних розразунків у RStudio.

2.4. Підготовка VCF-архіву для аналізу

Для полегшення та прискорення програмного аналізу, вихідний VCF файл було поділено на 6 частин відповідно до хромосомної приналежності, а саме: 2L, 2R, 3L, 3R, X, 4. (Рис. 2.2)

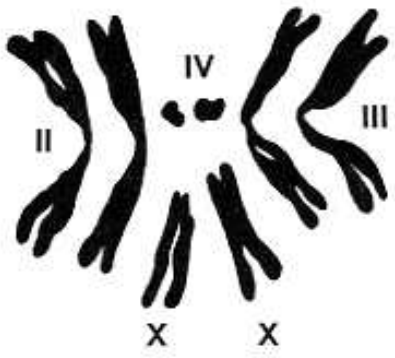

Рис. 2.2. Схема хромосом *D. melanogaster*

Дана процедура була виконана за допомогою програмного пакету SnpSift, що імплементований мовою програмування Java [41]. Для поділу за хромосомами використовувалась команда “SnpSift filter”. Повна команда виглядала так:

$ java -jar SnpSift.jar filter (CHROM=’2R’) droseu14.vcf > 2R.vcf

У дужках до параметру CHROM прирівнювався номер хромосоми, яку збирались відфільтрувати з архіву.

2.5. Побудова лінійних регресій в RStudio

З метою детекції клінальної мінливості поліморфізмів були побудовані генералізовані лінійні моделі, імплементовані мовою програмування R 3.6 за допомогою пакета vcfR [30, додаток 2]. Моделі також передбачали у своїй біноміальну похибку із вагою, пропорційною

до глибини покриття у кожному сайті та кількістю зібраних хромосом для кожної популяції.

Похибка була вирахувана за формулою (2.1):

$$□_{□□□} = (□_{□□□} * □_{□□} - 1)/(□_{□□□} + □_{□□}) \quad (2.1)$$

де $□_{□□□}$ позначає кількість відібраних хромосом у популяції, $□_{□□}$ - кількість ефективних прочитань (рідів) у кожному сайті. Для визначення клінальних поліморфізмів, широта, довгота, та висота над рівнем моря були регресовані згідно наступних моделей регресії:

$$□_{□} = □□□ + □_{□} \quad (2.2)$$

$$□_{□} = □□□□ + □_{□} \quad (2.3)$$

$$□_{□} = □□□ + □_{□} \quad (2.4)$$

де $□_{□}$ це частота алеля в і-му сайті геному та $\varepsilon_{□}$ означає біноміальну похибку виходячи з $□_{□□□}$.

Значення R-квадрат з кожної моделі лінійної регресії були інтерпретовані як метрика кореляції. Для корекції p-значень лінійних моделей шляхом мінімізації проблеми множинних порівнянь був використаний метод Хольма-Бонферроні.

Отримані результати були відфільтровані для отримання сайтів, що демонструють R-квадрат вище 0.1 як восени, так і влітку із p-значеннями не більше 0.05.

2.6. Пошук генів та функціональних взаємозв'язків за отриманими сайтами

Використовуючи бібліотеки Pandas та Allel у мові програмування Python, була екстрагована анотація сайтів, що представляють інтерес для

отримання генів, що продемонстрували значущу кореляцію (додаток 3). Задля отримання інформації про функціональні зв’язки між екстрагованими генами використовували базу даних білок-білкових взаємодій STRING [31].

РОЗДІЛ 3
РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ ТА ЇХНЄ ОБГОВОРЕННЯ

3.1. Побудова моделей лінійної регресії у сайтах геному *D. melanogaster*

Після побудови моделей лінійної регресії у кожному сайті геному D. melanogaster для визначення кореляції по географічній широті, довготі та висоті над рівнем моря було задетектовано 4 сайти (таблиця 3.1), що корелюють по довготі із задоволенням усіх поставлених вимог:

- R-квадрат >= 0.1 як в осінніх так і літніх популяціях
- p-значення R-квадрату після корекції на множинне тестування <= 0.05

Сайт	Ген	Хромосома
X_12308729	*Ten-a*	X
X_16726169	*CG32572*	X
4_623854	*mav*	4
4_808836	*Sox102F*	4

Таблиця 3.1. Сайти що проявляють клінальну мінливість та їх генна і хромосомна приналежність

Жодної кореляції по широті чи висоті не було зафіксовано, що відрізняється від даних досліджень в Австралії чи Північній Америці.

На графіках нижче представлені моделі лінійних регресій частот алелів у детектованих сайтах

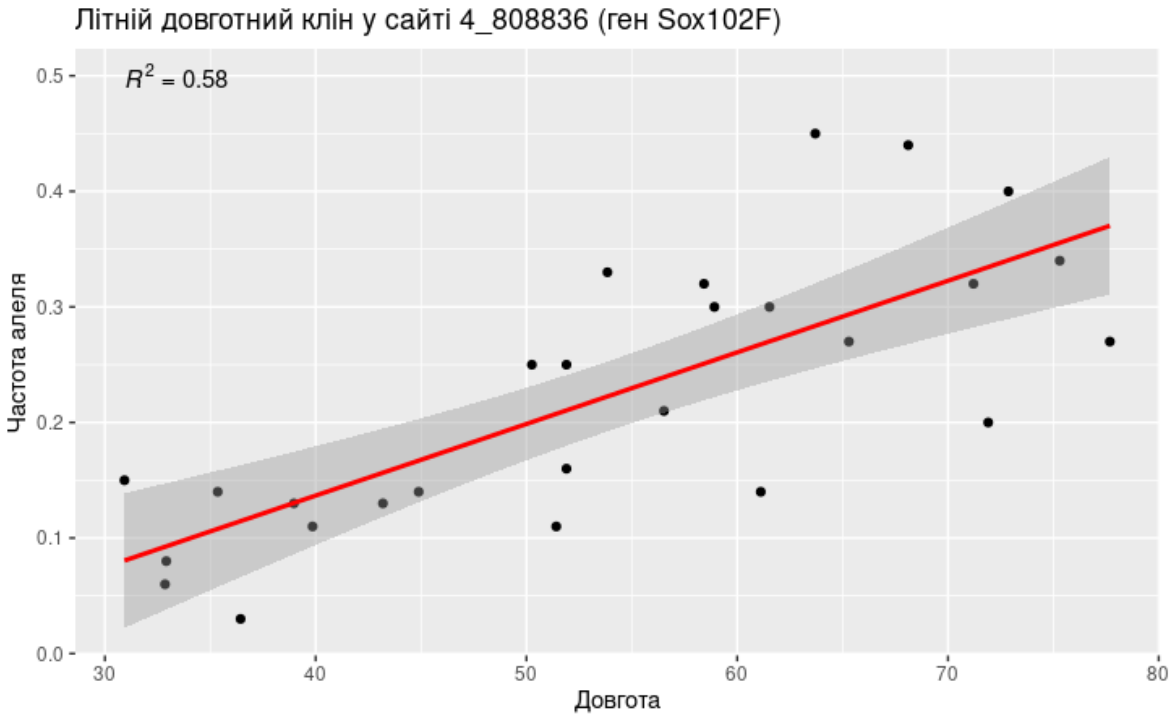

Рис. 3.1. Клінальність частоти алелю гену *Sox102F* влітку

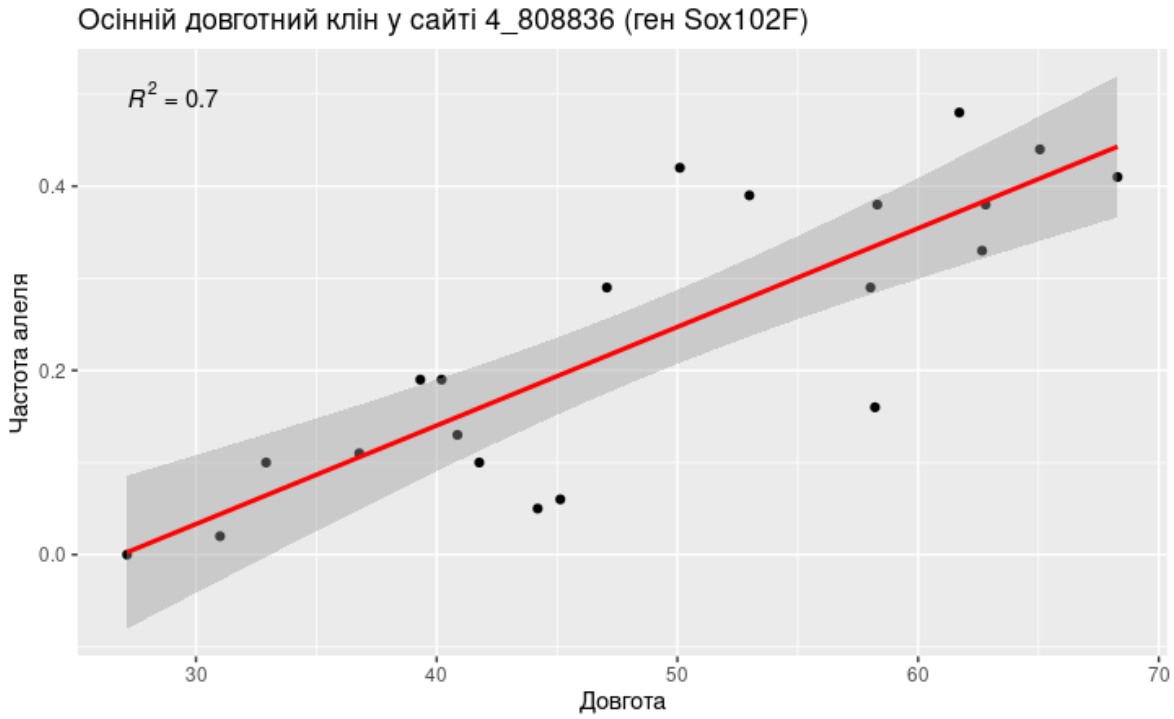

Рис. 3.2. Клінальність частоти алелю гену *Sox102F* восени

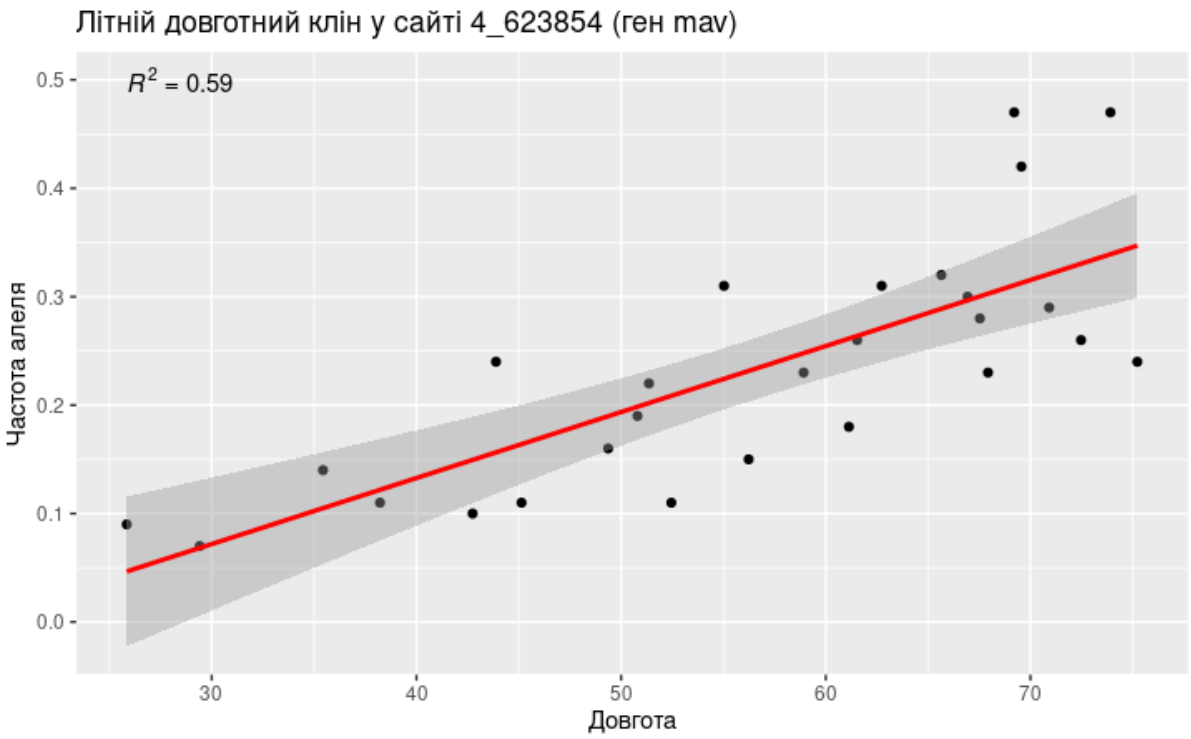

Рис. 3.3. Клінальність частоти алелю гену *mav* влітку

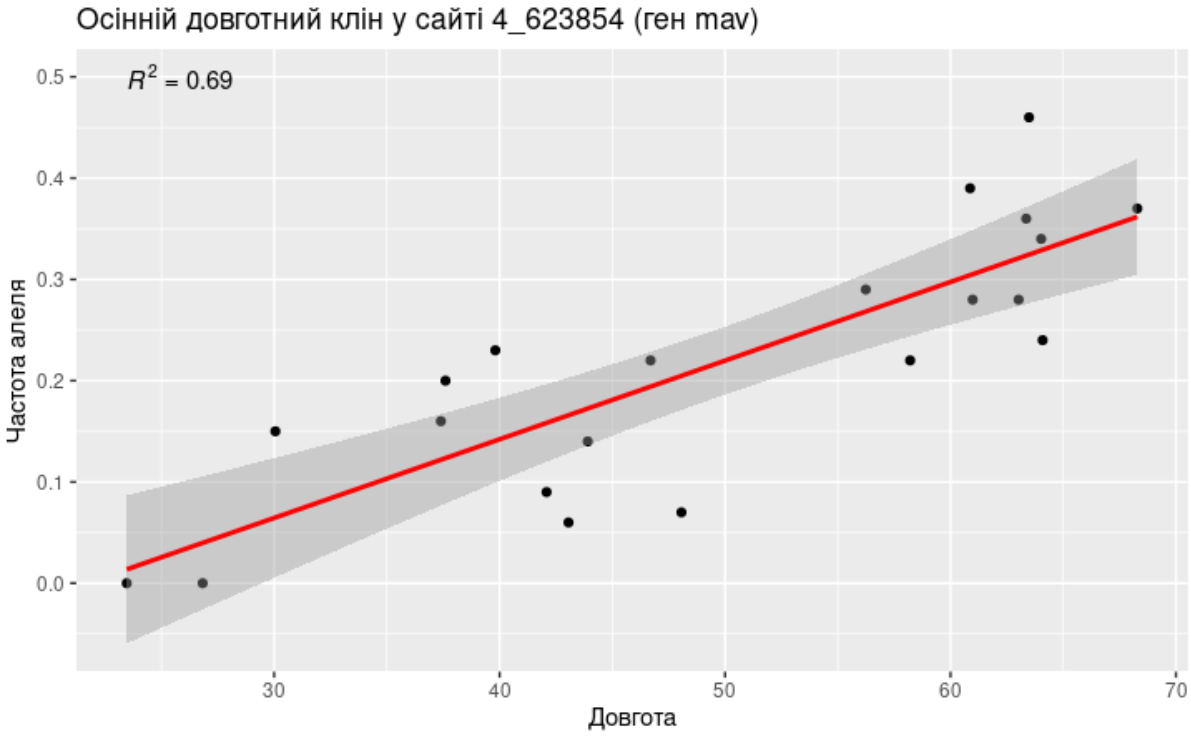

Рис. 3.4. Клінальність частоти алелю гену *mav* восени

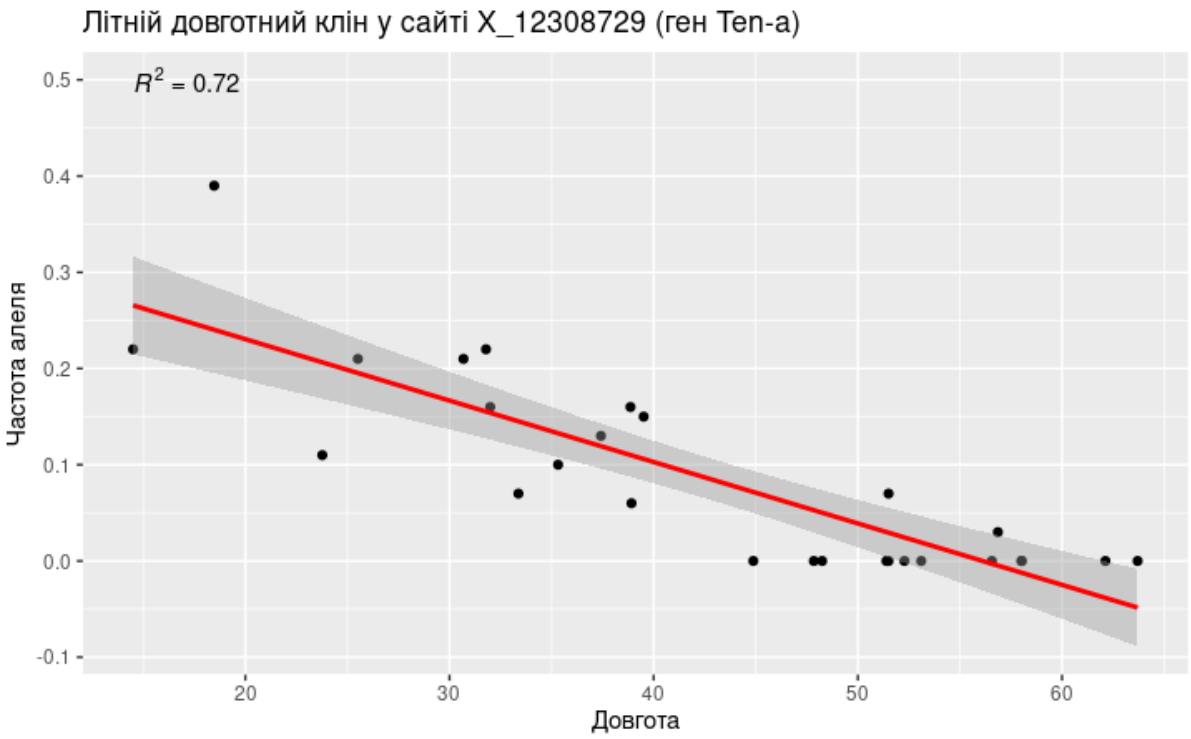

Рис. 3.5. Клінальність частоти алелю гену *Ten-a* влітку

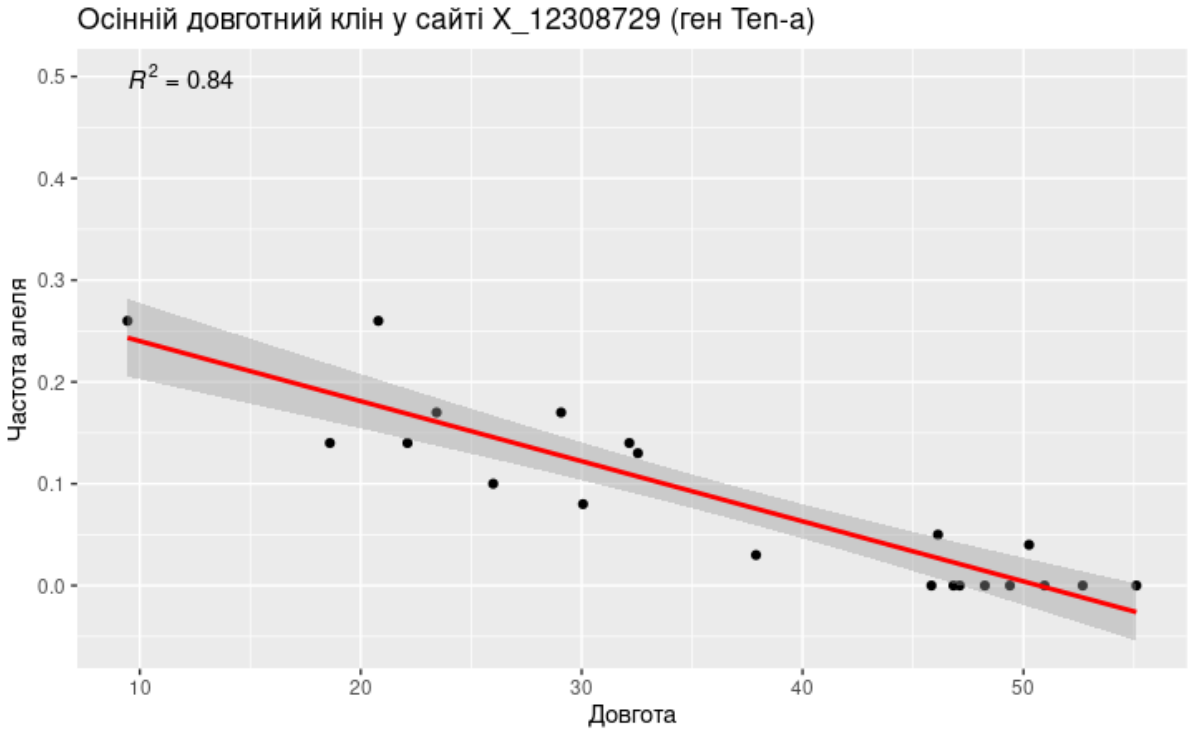

Рис. 3.6. Клінальність частоти алелю гену *Ten-a* восени

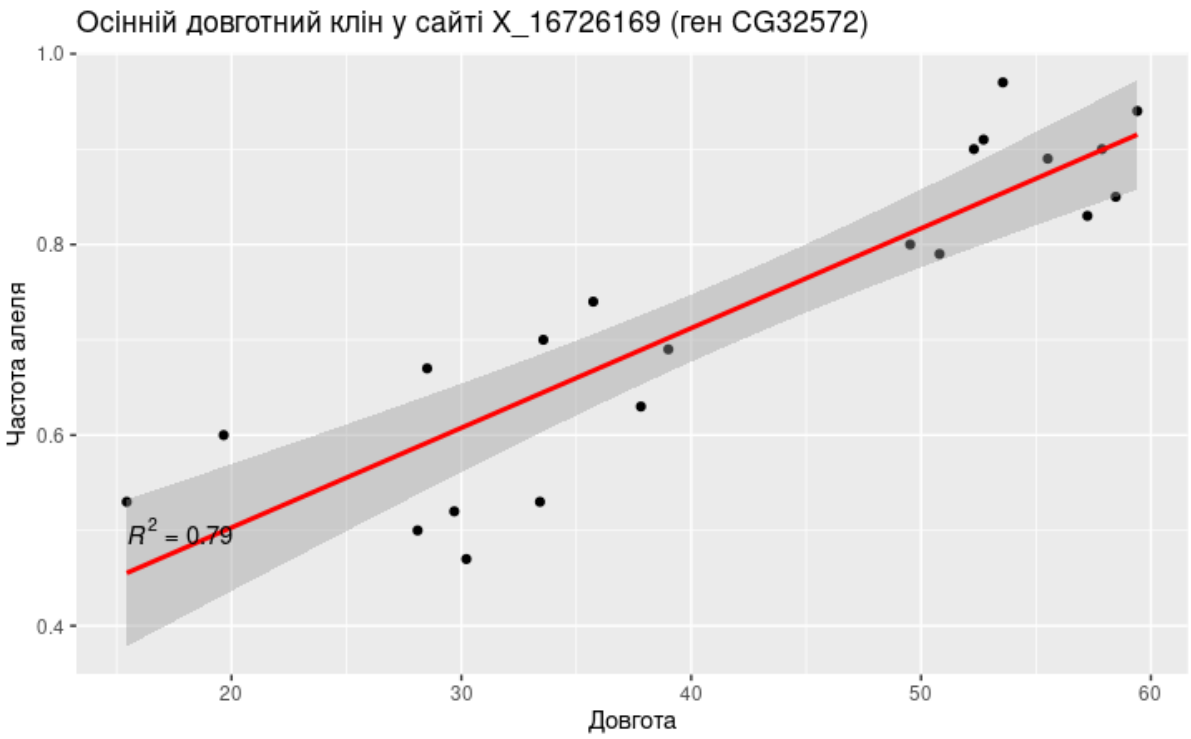

Рис. 3.7. Клінальність частоти алелю гену *CG32572* влітку

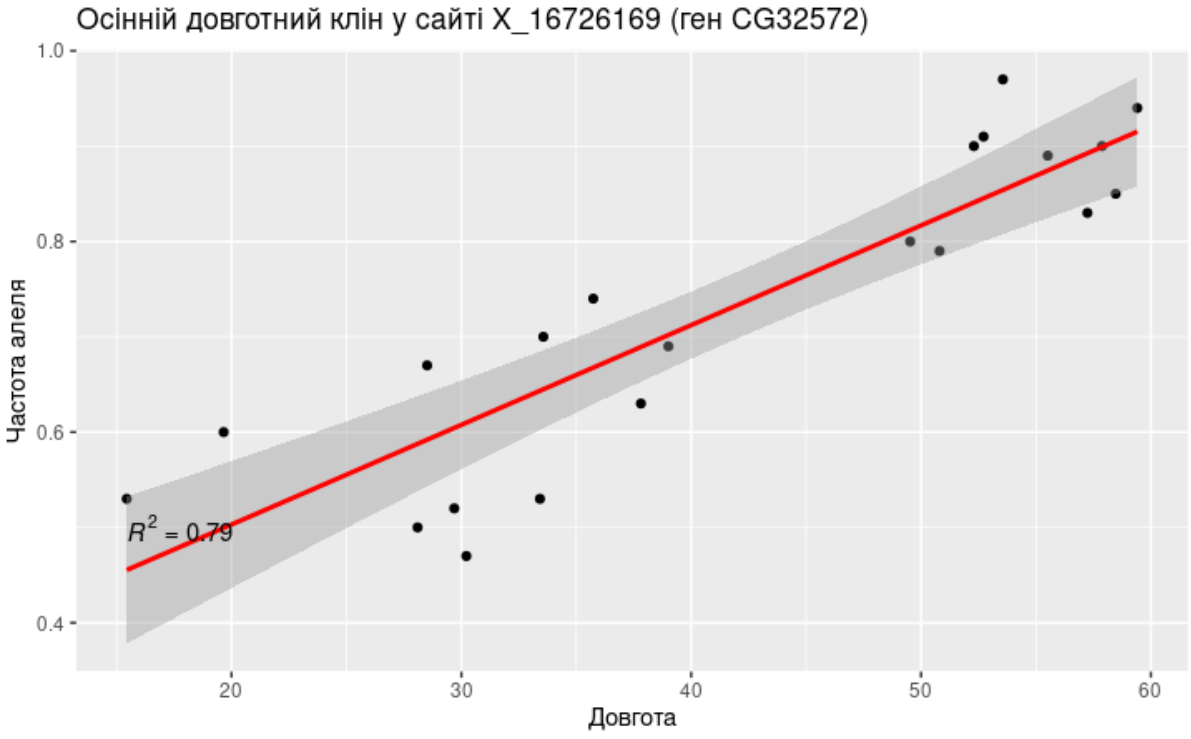

Рис. 3.8. Клінальність частоти алелю гену *CG32572* восени

3.2. Характеристика генів, що демонструють довготну клінальність

3.2.1. Ген *CG32572*

Ген *CG32572* експресується в черевному нервовому ланцюзі личинки та сперматозооні. Про властивості цього гена відомостей немає, утім у дослідженні Edwards et al вивчався вплив мутацій зокрема в *CG32572* на агресивну поведінку *D. melanogaster* [31]. Показано, що вставка P-елемента, транспозона дрозофіли, у послідовність *CG32572* призводила до підвищення рівня агресії в дрозофіл. Мутанти *CG32572* демонстрували підвищений рівень транскриптів гену. Крім цього, в особинах з мутацією *CG32572* було зафіксовано відсутність бета-долі у грибовидному тілі - аналозі мозку *D. melanogaster*.

3.2.2. Ген *mav*

Ген *mav* або maverick кодує білок, що залучений у TGF-β-сигнальному каскаді. Класичним індикатором сигнального каскаду TGF-β є Mad фосфорилювання. У личинкових нервово-м’язевих синапсах, Mad фосфорилювання служить вірним знаком TGF-β активації як ядрах моторних нейронів, так і в терміналах аксонів. В терміналах аксонів, P-Mad зібрані у вигляді окремих точок. Експериментально доведено, що зменшення експресії *Mav* у периферальній глії з двома різними *Mav*-РНКі конструктами повністю пригнічує або мінімізує імунореактивність P-Mad у синаптичних сайтах [32].

Утім, у випадку експресії *Mav*-РНКі в м'язах або моторних нейронах, різниці в синаптичних рівнях P-Mad не спостерігалось. Таким чином можна припустити, що *Mav* необхідний виключно у глії для активації TGF-β каскаду у синаптичних сайтах. Доказом цього, стала гіперекспресія *Mav*

у глії, що викликала підвищення рівня P-Mad сигналу в нейро-м’язевих синапсах. На відміну від глії, супресія *Mav* у моторних нейронах або м’язах суттєво не впливає на розмір нейро-м’язевих синапсів [32].

3.2.3. Ген *Sox102F*

Ген *Sox102F* кодує транскрипційний фактор з родини Sox, що регулює сигнальний каскад Wnt. Ролі цього каскаду включають розвиток серця та виконання серцево-судинних функцій, а також розвиток судин в крильцях. Крім цього, даний каскад контролює β-катенін, який провокує експресію ряду генів, приєднуючись до ядерної ДНК. У дослідженні Li et al [33] було з’ясовано, що сайленсинг *Sox102F* призводить до гострої серцевої дисфункції та структурних дефектів у дорослих мух, таких як серцева гіпертрофія, збільшений розмір серцевих камер, що супроводжуються пошкодженою будовою міофібрил. Також експресія *Sox102F* може регулювати безкрилу (wg) експресію.

3.2.4. Ген *Ten-a*

Ген *Ten-a* - тенеурин, кодує трансмембранний білок II типу, з зовнішньоклітинним доменом, що опосередковує гомофільні взаємодії з самим собою та продуктом експресії *Ten-m.* Цей білок також взаємодіє з білками регуляції цитоскелету. Ten-a регулює спрямування аксонів у ембріональній нервовій системі, сполучення синаптичних партнерів, а також організацію синапсів у нюхових та нейром’язевих системах. Для дослідження функції Ten-a у роботі Mosca et al проаналізували особини з нульовим алелем Ten-a та личинки РНК-інтерференцією гену у нейронах чи м’язах [34]. Такі маніпуляції вплинули на кількість аксонних терміналів: кількість зменшилась на 55% та помітно зросла поява великих

аксонних терміналів. Обидва явища свідчать про пошкоджений синаптичний морфогенез. В Ten-a мутантах, морфогенез синаптичних відновився після поновлення експресії Ten-a у нейронах, утім не у м'язах [34]. Крім цього, зафіксовано, що в Ten-a мутантах амплітуда збуджуючого постсинаптичного потенціалу була знижена на 28%. Спонтанні мініатюрні збуджуючі постсинаптичні потенціали демонструють зниження на 20% в амплітуді та 46% у частоті [34].

3.3. Пошук функціональних взаємозв'язків між генами

За допомогою бази даних STRING був здійснений скринінг генів *CG32572*, *Sox102F*, *Ten-a* та *mav* на предмет наявності метаболічних взаємозв'язків у *D. melanogaster*. Згідно результатів, жодних зв'язків чи взаємозалежностей між цими генами та білками, що вони експресують немає (рис. 3.9).

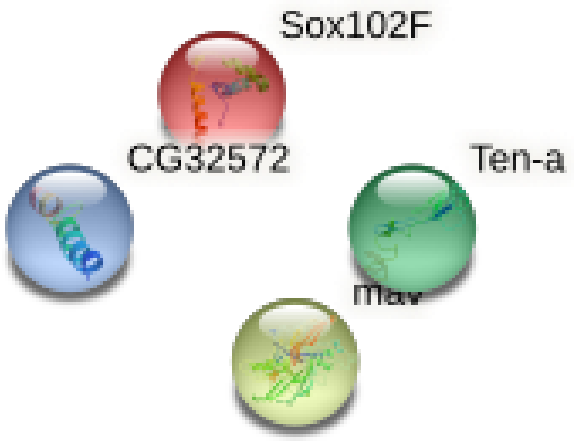

Рис. 3.9. Мапа зв'язків між генами *D. melanogaster* з довготною клінальністю

Таким чином, причина, чому саме ці 4 гени єдиними проявили довготну клінальність у популяціях Європи підлягає подальшому

вивченню. Утім, причетність генів *mav* та *Ten-a до розвитку та* функціонування нервової системи у *D. melanogaster* дає підстави припускати, що існує певна залежність, що досі фактично не зафіксована.

Також, враховуючи, що гени з клінальністю можно умовно поділити на пари: *Sox102F* та *mav* на 4 хромосомі та *Ten-a* і *CG32572* на X хромосомі, можливе зчеплення цих генів.

ВИСНОВКИ

1. Було виявлено 4 сайти зі статистично значущою кореляцією по довготі, що належать наступним генам: *Ten-a*, *CG32572* (X-хромосома), *Sox102F*, *mav* (4 хромосома).

2. *Ten-a* залучений до організації синаптичних зв'язків; *Sox102F* бере участь у сигнальному каскаді Wnt; *mav* кодує білок, залучений у сигнальному каскаді TGF-β; властивості *CG32572* досі достеменно невідомі.

3. Кореляцій частот алелів по географічній широті чи висоті над рівнем моря не виявлено.

4. Перераховані гени не демонструють молекулярних взаємозв'язків згідно базі даних STRING та підлягають подальшому вивченню.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Endler, J. A. Geographic variation, speciation, and clines. *Monographs in population biology* vol. 10 (1977): 1-246.

2. Endler, J. A. Natural Selection in the Wild.(MPB-21), Volume 21. *Natural Selection in the Wild.(MPB-21), Volume 21*. Princeton University Press, 2020.

3. Lachaise, D. Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evolutionary biology*. Springer, Boston, MA, 1988. 159-225.

4. Keller, Andreas. Drosophila melanogaster's history as a human commensal. *Current biology: CB vol. 17,3* (2007): R77-81.

5. David, Jean R., and Pierre Capy. Genetic variation of Drosophila melanogaster natural populations. *Trends in Genetics* 4.4 (1988): 106-111.

6. Kohler, Robert E. *Lords of the fly: Drosophila genetics and the experimental life*. University of Chicago Press, 1994.

7. Kohler, Robert E. *Lords of the fly: Drosophila genetics and the experimental life*. University of Chicago Press, 1994.

8. Adams, Mark D., et al. "The genome sequence of Drosophila melanogaster." *Science* 287.5461 (2000): 2185-2195.

9. Wright, Sewall. "Dobzhansky's Genetics of Natural Populations, I-XLIII." (1982): 1102-1106.

10. Hudson, R R. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology* vol. 23,2 (1983): 183-201.

11. Langley, C. Restriction map variation in the Adh region of Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* vol. 79,18 (1982): 5631-5.

12. Aquadro, C F et al. Molecular population genetics of the alcohol dehydrogenase gene region of Drosophila melanogaster. *Genetics* vol. 114,4 (1986): 1165-90.

13. Powell, Jeffrey R. *Progress and prospects in evolutionary biology: the Drosophila model*. Oxford University Press, 1997.

14. Charlesworth, Brian, and D. Charlesworth. Population genetics from 1966 to 2016. *Heredity* 118.1 (2017): 2-9.

15. Hudson, R. A test of neutral molecular evolution based on nucleotide data. *Genetics* vol. 116,1 (1987): 153-9.

16. McDonald, J.H., Kreitman, M.. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* vol. 351,6328 (1991): 652-4.

17. Booker, Tom R et al. Detecting positive selection in the genome. *BMC biology* vol. 15,1 98. 30 Oct. 2017,

18. Kern, Andrew D, and Matthew W Hahn. The Neutral Theory in Light of Natural Selection. *Molecular biology and evolution* vol. 35,6 (2018): 1366-1371.

19. Begun, D., Aquadro, C. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* vol. 356,6369 (1992): 519-20.

20. Bergland, A. (2014) Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *PLoS genetics* vol. 10,11.

21. Behrman, E. L., et al. Seasonal variation in life history traits in two Drosophila species. *Journal of evolutionary biology* 28.9 (2015): 1691-1704.

22. St Johnston, Daniel. The art and design of genetic screens: Drosophila melanogaster. *Nature reviews. Genetics* vol. 3,3 (2002): 176-88.

23. Hoskins, R. A. The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome research* vol. 25,3 (2015): 445-58.

24. Sessegolo, Camille et al. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biology letters* vol. 12,8 (2016): 20160407.

25. Drosophila 12 Genomes Consortium et al. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* vol. 450,7167 (2007): 203-18.

26. Celniker, Susan E et al. “Unlocking the secrets of the genome. *Nature* vol. 459,7249 (2009): 927-30.

27. Sotka, E. E. (2008). *Clines. Encyclopedia of Ecology, 613–618.*

28. Martin Kapun, Daniel K. Fabian, Jérôme Goudet, Thomas Flatt, Genomic Evidence for Adaptive Inversion Clines in *Drosophila melanogaster*, *Molecular Biology and Evolution*, Volume 33, Issue 5, May 2016, Pages 1317–1336.

29. Bergland, A. O., Behrman, E. L., O’Brien, K. R., Schmidt, P. S., & Petrov, D. A. (2014). Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in *Drosophila. PLoS Genetics,* 10(11).

30. https://www.r-project.org/

31. Edwards, Alexis C et al. Mutations in many genes affect aggressive behavior in Drosophila melanogaster. *BMC biology* vol. 7 29. 11 Jun. 2009.

32. Fuentes-Medel, Yuly et al. Integration of a retrograde signal during synapse formation by glia-secreted TGF-β ligand. *Current biology : CB* vol. 22,19 (2012): 1831-8.

33. Li, Airong et al. Silencing of the Drosophila ortholog of SOX5 in heart leads to cardiac dysfunction as detected by optical coherence tomography. *Human molecular genetics* vol. 22,18 (2013): 3798-806.

34. Mosca, Timothy, J. (2012) Trans-synaptic Teneurin signalling in neuromuscular synapse organization and target choice. *Nature* vol. 484,7393 237-41.

35. Adrion, Jeffrey R et al. Revisiting classic clines in Drosophila melanogaster in the age of genomics. *Trends in genetics : TIG* vol. 31,8 (2015): 434-44.

36. https://colab.research.google.com/

37. https://cran.r-project.org/web/packages/vcfR/vignettes/intro_to_vcfR.html\

38. https://scikit-allel.readthedocs.io/en/stable/#

39. Charlesworth, Deborah et al. The sources of adaptive variation. *Proceedings. Biological sciences* vol. 284,1855 (2017): 20162864. doi:10.1098/rspb.2016.2864

40. Barton, N H. Clines in polygenic traits. *Genetical research* vol. 74,3 (1999): 223-36. doi:10.1017/s001667239900422x

41. http://pcingola.github.io/SnpEff/ss_introduction/

42. Vigue, C.L., Johnson, F.M. Isozyme variability in species of the genus *Drosophila*. VI. Frequency-property-environment relationships of allelic alcohol dehydrogenases in *D. melanogaster* . *Biochem Genet* 9, 213–227 (1973).

43. Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. *New England Journal of Medicine,* 358(3), 252–260.

44. Karan, D., & Parkash, R. (1998). Desiccation tolerance and starvation resistance exhibit opposite latitudinal clines in Indian geographical populations of *Drosophila kikkawai*. *Ecological Entomology,* 23(4), 391–396.

45. Duchen, Pablo, et al. Demographic inference reveals African and European admixture in the North American Drosophila melanogaster population. *Genetics* 193.1 (2013): 291-301.

46. Hey, Jody, and Richard M. Kliman. Population genetics and phylogenetics of DNA sequence variation at multiple loci within the Drosophila melanogaster species complex. *Molecular biology and evolution* 10.4 (1993): 804-822.

47. Cariou, Marie Louise. Biochemical phylogeny of the eight species in the *Drosophila melanogaster* subgroup, including *D. sechellia* and *D. orena*. *Genetics Research* 50.3 (1987): 181-185.

48. David, J. R., et al. "Thoracic trident pigmentation in Drosophila melanogaster: differentiation of geographical populations." *Génétique sélection évolution* 17.2 (1985): 211-224.

49. McKenzie, A., Parsons, P.A. (1974) The Genetic Architecture of Resistance to Desiccation in Populations of Drosophila melanogaster and D. simulans. *Australian Journal of Biological Sciences* 27, 441-456.

50. Harshman, L.G., Hoffmann, A.A., Clark, A.G. (1999). Selection for starvation resistance in Drosophila melanogaster: physiological correlates, enzyme activities and multiple stress responses. *J. Evol. Biol.* 12(2): 370--79.

51. Schmidt, P. S., & Paaby, A. B. (2008). Reproductive diapause and life-history clines in North American populations of *Drosophila Melanogaster. Evolution, 62(5),* 1204–1215.

52. Fabian, D. K., Kapun, M., Nolte, V., Kofler, R., Schmidt, P. S., Schlötterer, C., & Flatt, T. (2012). Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology, 21(19),* 4748–4769.

53. Hoffmann, Ary A, and Andrew R Weeks. Climatic selection on genes and traits after a 100 year-old invasion: a critical look at the temperate-tropical clines in *Drosophila melanogaster* from eastern Australia. *Genetica vol. 129,2* (2007): 133-47.

54. Bergland, A.O., Tobler, R., González, J., Schmidt, P. and Petrov, D. (2016), Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. *Mol Ecol,* 25: 1157-1174.

55. Aulard, S., Vaudin, P., Ladeveze, V., Chaminade, N., Periquet, G., Lemeunier, F. (2004). Maintenance of a large pericentric inversion generated by the hobo transposable element in a transgenic line of *Drosophila melanogaster. Heredity 92(3):* 151-155.

56. Fabian, D. K et al. Spatially varying selection shapes life history clines among populations of *Drosophila melanogaster* from sub-Saharan Africa. *Journal of evolutionary biology vol. 28,4* (2015): 826-40.

57. Sambucetti, P., Loeschke, V. and Norry, F.M. (2006), Developmental time and size-related traits in Drosophila buzzatii along an altitudinal gradient from Argentina. *Hereditas, 143:* 77-83.

58. Carson, H. L., & Stalker, H. D. (1947). Gene Arrangements in Natural Populations of Drosophila robusta Sturtevant. *Evolution, 1(3),* 113

59. Prakash, S., Caldwell, J.C., Eberl, D.F., Clandinin, T.R. (2005). *Drosophila* N-cadherin mediates an attractive interaction between photoreceptor axons and their targets. *Nat. Neurosci. 8(4):* 443-450.

60. Dillon, M. E. (2006). *Drosophila melanogaster* locomotion in cold thin air. *Journal of Experimental Biology, 209(2),* 364–371.

61. Körner, C. (2007). The use of “altitude” in ecological research. *Trends in Ecology & Evolution, 22(11),* 569574.

62. Blanckenhorn, W. U. (1997). Altitudinal life history variation in the dung flies *Scathophaga stercoraria* and *Sepsis cynipsea. Oecologia, 109(3),* 342–352.

63. Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in ecology & evolution vol. 29,1* (2014): 51-63.

64. Schlötterer, C. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature reviews. Genetics vol. 15,11* (2014)*:* 749-63.

65. Wang, Y., Hey, J. Estimating Divergence Parameters With Small Samples From a Large Number of Loci. *Genetics, vol. 184, Issue 2,* (2010): 363379.

66. Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genetics, 6(2)*

67. Van Tassell, Curtis P et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature methods vol. 5,3 (2008)*: 247-52.

Додаток 1. Таблиця географічних даних популяцій *Drosophila melanogaster* Європи

ID	Country	Location	Date	Number ID	Latitude	Longitude	Altitude	Season	flies pooled	chrom	n
AT_Mau_14_01	Austria	Mauternbach	2014-07-20	1	48.375	15.560	172	S	40	0	80
AT_Mau_14_02	Austria	Mauternbach	2014-10-19	2	48.375	15.560	172	S	40	0	80
TR_Yes_14_03	Turkey	Yesiloz	2014-08-31	3	40.231	32.260	1380	S	40	0	80
TR_Yes_14_04	Turkey	Yesiloz	2014-10-23	4	40.231	32.260	1380	S	40	0	80
FR_Vil_14_05	France	Viltain	2014-08-18	5	48.754	2.158	153	S	40	0	80
FR_Vil_14_07	France	Viltain	2014-10-27	7	48.754	2.158	153	S	40	0	80
FR_Got_14_08	France	Gotheron	2014-07-08	8	44.977	4.930	181	S	40	0	80
UK_She_14_09	United Kingdom	Sheffield	2014-08-25	9	53.391	-1.521	100	S	40	0	80
UK_Sou_14_10	United Kingdom	South Queensferry	2014-07-14	10	55.973	-3.351	9	S	40	0	80
CY_Nic_14_11	Cyprus	Nicosia	2014-08-10	11	35.066	33.323	163	S	40	0	80

UK_Mar_14_12	United Kingdom	Market Harborough	2014-10-20	2	1	52.476	0.922	-0	8	0	4	0	8
UK_Lut_14_13	United Kingdom	Lutterworth	2014-10-20	3	1	52.434	1.103	-26	1	0	4	0	8
DE_Bro_14_14	Germany	Brogingen	2014-06-26	4	1	48.218	8.820	73	1	0	4	0	8
DE_Bro_14_15	Germany	Brogingen	2014-10-15	5	1	48.218	8.820	73	1	0	4	0	8
UA_Yal_14_16	Ukraine	Yalta	2014-06-20	6	1	44.502	34.166	2	7	0	4	0	8
UA_Yal_14_18	Ukraine	Yalta	2014-08-27	8	1	44.502	34.166	2	7	0	4	0	8
UA_Ode_14_19	Ukraine	Odesa	2014-07-03	9	1	46.442	30.772	4	5	0	4	0	8
UA_Ode_14_20	Ukraine	Odesa	2014-07-22	0	2	46.442	30.772	4	5	0	4	0	8
UA_Ode_14_21	Ukraine	Odesa	2014-08-29	1	2	46.442	30.772	4	5	0	4	0	8
UA_Ode_14_22	Ukraine	Odesa	2014-10-10	2	2	46.442	30.772	4	5	0	4	0	8
UA_Kyi_14_23	Ukraine	Kyiv	2014-08-09	3	2	50.344	30.489	79	1	0	4	0	8

UA_Kyi_14_24	Ukraine	Kyiv	2014-09-08	4	2	0.344	0.489	3	79	1	0	4	0	8
UA_Var_14_25	Ukraine	Varva	2014-08-18	5	2	0.485	2.714	3	25	1	0	4	0	8
UA_Pyr_14_26	Ukraine	Pyriatyn	2014-08-20	6	2	0.254	2.519	3	14	1	0	4	0	8
UA_Dro_14_27	Ukraine	Drogobych	2014-08-24	7	2	9.330	3.503	2	75	2	0	4	0	8
UA_Cho_14_28	Ukraine	Chornobyl	2014-09-13	8	2	1.373	0.138	3	21	1	0	4	0	8
UA_Cho_14_29	Ukraine	Chornobyl Yaniv	2014-09-13	9	2	1.387	0.073	3	21	1	0	4	0	8
SE_Lun_14_30	Sweden	Lund	2014-07-31	0	3	5.694	3.198	1	1	5	0	4	0	8
DE_Mun_14_31	Germany	Munich	2014-06-19	1	3	8.180	1.610	1	20	5	0	4	0	8
DE_Mun_14_32	Germany	Munich	2014-09-03	2	3	8.180	1.610	1	20	5	0	4	0	8
PT_Rec_14_33	Portugal	Recarei	2014-09-26	3	3	1.150	8.410	-	75	1	0	4	0	8
ES_Gim_14_34	Spain	Gimenells (Lleida)	2014-10-20	4	3	1.618	0.620	0	73	1	0	4	0	8
ES_Gim_14	Spain	Gimenells	2014-	5	3	1.618	0.620	0	73	1	0	4	0	8

_35		(Lleida)	08-13									
FI_Aka_14_36	Finland	Akaa	2014-07-25	6	3	1.100	3.520	28		0	0	8
FI_Aka_14_37	Finland	Akaa	2014-08-27	7	3	1.100	3.520	28		0	0	8
FI_Ves_14_38	Finland	Vesanto	2014-07-26	8	3	2.550	6.240	21		3	6	6
DK_Kar_14_39	Denmark	Karensminde	2014-09-01	9	3	5.945	0.213	15		0	0	8
DK_Kar_14_41	Denmark	Karensminde	2014-11-25	1	4	5.945	0.213	15		0	0	8
CH_Cha_14_42	Switzerland	Chalet à Gobet	2014-07-24	2	4	6.567	.702	72		0	0	8
CH_Cha_14_43	Switzerland	Chalet à Gobet	2014-10-05	3	4	6.567	.702	72		0	0	8
AT_See_14_44	Austria	Seeboden	2014-08-17	4	4	6.814	3.508	91		0	0	8
UA_Kha_14_45	Ukraine	Kharkiv	2014-07-26	5	4	9.819	6.055	41		0	0	8
UA_Kha_14_46	Ukraine	Kharkiv	2014-09-14	6	4	9.819	6.055	41		0	0	8
UA_Cho_14_47	Ukraine	Chornobyl	2014-09-13	7	4	1.273	0.221	21		0	0	8

		Applegarden											
UA_Cho_14_48	Ukraine	Chornobyl Polisske	2014-09-13	48	4	1.279	9.394	21	1	5	3	0	7
UA_Kyi_14_49	Ukraine	Kyiv	2014-10-11	49	4	0.344	0.489	79	1	0	4	0	8
UA_Uma_14_50	Ukraine	Uman	2014-10-01	50	5	8.753	0.206	14	2	0	4	0	8
RU_Val_14_51	Russia	Valday	2014-08-17	51	5	7.979	3.244	17	2	0	4	0	8

Додаток 2. Програма мовою R для побудови та запису моделей лінійної регресії.

```
library(vcfR)
library(dplyr)
library(readxl)
```

```
library(xlsx)
library(tibble)
library(sys)

start <- Sys.time()
#читає таблицю з популяціями 2014-16
excel <- read_excel("population_seasons.xlsx")

#отримує з таблиці широту, довготу та висоту
latitude <- excel$Latitude
longitude <- excel$Longitude
altitude <- excel$Altitude
season <- excel$Season
nflies <- excel$nflies
nchrom <- excel$nchrom

vcf <- read.vcfR("chr4.vcf")
#виділяє з архіву алельні частоти популяцій
freq <- extract.gt(vcf, element = c('FREQ'), as.numeric = TRUE)
#reads <- extract.gt(vcf, element = c('RD'), as.numeric = TRUE)
reads_num <- extract.gt(vcf, element = c('DP'), as.numeric = TRUE)

names <- freq[,0]

# створює вектори для запису в них значень R-squared
summer_rsq_lat <- vector(length = nrow(freq))
autumn_rsq_lat <- vector(length = nrow(freq))
summer_rsq_long <- vector(length = nrow(freq))
autumn_rsq_long <- vector(length = nrow(freq))
summer_rsq_alt <- vector(length = nrow(freq))
autumn_rsq_alt <- vector(length = nrow(freq))

#створює вектори для запису p-value
summer_p_lat <- vector(length = nrow(freq))
autumn_p_lat <- vector(length = nrow(freq))
summer_p_long <- vector(length = nrow(freq))
autumn_p_long <- vector(length = nrow(freq))
summer_p_alt <- vector(length = nrow(freq))
autumn_p_alt <- vector(length = nrow(freq))

#цикл для проходження по усім рядкам (сайтам) гену;
#побудови лінійної регресії та збереження R-squared і p-value
for (i in 1:nrow(freq))
```

```
        {
          row <- freq[i, ]
          #rd = read[i, ]
          nrd <- reads_num[i, ]

          neff <- (nchrom * nrd - 1)/(nchrom + nrd)

          latitude_corrected <- latitude + neff
          longitude_corrected <- longitude + neff
          altitude_corrected <- altitude + neff

          sample_lat <- data.frame(latitude_corrected, row, season)
          summer_sample_lat <- filter(sample_lat, season == "S")
          autumn_sample_lat <- filter(sample_lat, season == "F")

          sample_long <- data.frame(longitude_corrected, row, season)
          summer_sample_long <- filter(sample_long, season == "S")
          autumn_sample_long <- filter(sample_long, season == "F")

          sample_alt <- data.frame(altitude_corrected, row, season)
          summer_sample_alt <- filter(sample_alt, season == "S")
          autumn_sample_alt <- filter(sample_alt, season == "F")

          sum.regression_lat <- lm(row ~ latitude_corrected,
data=summer_sample_lat)
          aut.regression_lat <- lm(row ~ latitude_corrected,
data=autumn_sample_lat)

          sum.regression_long <- lm(row ~ longitude_corrected,
data=summer_sample_long)
          aut.regression_long <- lm(row ~ longitude_corrected,
data=autumn_sample_long)

          sum.regression_alt <- lm(row ~ altitude_corrected,
data=summer_sample_alt)
          aut.regression_alt <- lm(row ~ altitude_corrected,
data=autumn_sample_alt)

          summer_rsq_lat[i] = summary(sum.regression_lat)$r.squared
          autumn_rsq_lat[i] = summary(aut.regression_lat)$r.squared
          summer_p_lat[i] = summary(sum.regression_lat)$coefficients[2, 4]
          autumn_p_lat[i] = summary(aut.regression_lat)$coefficients[2, 4]
```

```
  summer_rsq_long[i] = summary(sum.regression_long)$r.squared
  autumn_rsq_long[i] = summary(aut.regression_long)$r.squared
  summer_p_long[i] = summary(sum.regression_long)$coefficients[2, 4]
  autumn_p_long[i] = summary(aut.regression_long)$coefficients[2, 4]

  summer_rsq_alt[i] = summary(sum.regression_alt)$r.squared
  autumn_rsq_alt[i] = summary(aut.regression_alt)$r.squared
  summer_p_alt[i] = summary(sum.regression_alt)$coefficients[2, 4]
  autumn_p_alt[i] = summary(aut.regression_alt)$coefficients[2, 4]
 }

 squares <- data.frame(summer_rsq_lat, autumn_rsq_lat,
summer_rsq_long, autumn_rsq_long, summer_rsq_alt,
                       autumn_rsq_alt)

 # squares <- data.frame(summer_rsq_long, autumn_rsq_long)

 for (i in colnames(squares)){
  squares[i] <- round(squares[i], digits = 3)
 }

 #summer_p_lat <- round(p.adjust(summer_p_lat , method="holm"), digits
= 3)
 #autumn_p_lat <- round(p.adjust(autumn_p_lat , method="holm"), digits
= 3)
 #summer_p_long <- round(p.adjust(summer_p_long , method="holm"),
digits = 3)
 #autumn_p_long <- round(p.adjust(autumn_p_long , method="holm"),
digits = 3)
 #summer_p_alt <- round(p.adjust(summer_p_alt , method="holm"), digits
= 3)
 #autumn_p_alt <- round(p.adjust(autumn_p_alt , method="holm"), digits
= 3)

 summer_p_lat <- round(summer_p_lat, digits = 3)
 autumn_p_lat <- round(autumn_p_lat, digits = 3)
 summer_p_long <- round(summer_p_long, digits = 3)
 autumn_p_long <- round(autumn_p_long, digits = 3)
 summer_p_alt <- round(summer_p_alt, digits = 3)
 autumn_p_alt <- round(autumn_p_alt, digits = 3)
```

```
        general_panel_lat <- data.frame(names, squares$summer_rsq_lat,
squares$autumn_rsq_lat,
                                        summer_p_lat, autumn_p_lat)
        general_panel_long <- data.frame(names, squares$summer_rsq_long,
squares$autumn_rsq_long,
                                         summer_p_long, autumn_p_long)
        general_panel_alt <- data.frame(names, squares$summer_rsq_alt,
squares$autumn_rsq_alt,
                                         summer_p_alt, autumn_p_alt)

        colnames(general_panel_lat) <- c("corr_sum_lat", "corr_aut_lat", "sum p-
val lat",
                                         "aut p-val lat")
        colnames(general_panel_long) <- c("corr_sum_long", "corr_aut_long",
"sum p-val long",
                                         "aut p-val long")
        colnames(general_panel_alt) <- c("corr_sum_alt", "corr_aut_alt", "sum p-
val alt",
                                         "aut p-val alt")

        #фільтрує сайти з R-squared >= 0.1 в популяціях обох сезонів
        filter_total_lat <- general_panel_lat %>%
          rownames_to_column('site') %>%
          filter_at(vars(starts_with("c")), all_vars(. >= 0.1)) %>%
          column_to_rownames('site')

        filter_total_long <- general_panel_long %>%
          rownames_to_column('site') %>%
          filter_at(vars(starts_with("c")), all_vars(. >= 0.1)) %>%
          column_to_rownames('site')

        filter_total_alt <- general_panel_alt %>%
           rownames_to_column('site') %>%
           filter_at(vars(starts_with("c")), all_vars(. >= 0.1)) %>%
           column_to_rownames('site')

        write.csv(filter_total_lat, "chr4_lat_final.csv")
        write.csv(filter_total_long, "chr4_long_final.csv")
        write.csv(filter_total_alt, "chr4_alt_final.csv")

        end <- Sys.time()
        duration <- end - start
        print(duration)
```

Додаток 3. Програма мовою Python для анотації до сайтів з клінальністю

```
import allel
import numpy as np
import pandas as pd

callset = allel.vcf_to_dataframe('X.vcf', fields='*')
new = callset.filter(['CHROM', 'POS', 'ANN'], axis=1)

r2 = pd.read_csv('X_regression_long_final.csv')
r2_filtered = r2[(r2['sum p-val long'] <= 0.05) & (r2['aut p-val long'] <= 0.05)]
print(r2_filtered)

buffer1 = dict()

# запис сайтів з r2_filtered
buffer1['4'] = [623854, 808836]
```

```
buffer1['X'] = [12308729, 16726169]

for key, value in buffer1.items():
    ex = new.loc[(new['CHROM'] == key) & (new['POS'].isin(value))]
    print(ex)
```