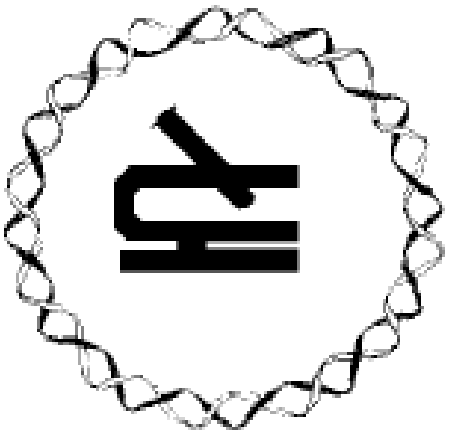
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ І. І. МЕЧНИКОВА
БІОЛОГІЧНИЙ ФАКУЛЬТЕТ

МЕТОДИЧНІ ВКАЗІВКИ

до практичної роботи з дисципліни «Біоінформатичний аналіз даних геномного секвенування» за розділом «Пошук кластерів генів вторинних метаболітів » для здобувачів третього (освітньо-наукового) рівня вищої освіти зі спеціальності 162 «Біотехнології та біоінженерія» та 091 «Біологія»



ОДЕСА
ОНУ
2023

УДК 579.6+ 578

Рекомендовано до друку

Вченою радою біологічного факультету
ОНУ імені І. І. Мечникова.
Протокол № ? від ??????? р.

Рецензенти:
Світлана МІРОСЬ, кандидат біологічних наук, доцент кафедри генетики ОНУ імені І. І. Мечникова;

Оксана ЗІНЧЕНКО, кандидат біологічних наук, доцент кафедри мікробіології, вірусології та біотехнології ОНУ імені І. І. Мечникова.

Пошук кластерів генів вторинних метаболітів: метод. вказівки до проведення практичних занять з курсу / Наталія ВАСИЛЬЄВА. – Одеса : Одес. нац. ун-т ім. І. І. Мечникова, 2023. – с.

Методичні вказівки до проведення практичних занять присвячені розгляду роботі з програмою antismash. Однією з основних завдань після отримання послідовностей після секвенування є перевірка їх якості. FastQC – розроблений в лабораторії Babraham Bioinformatics є дуже популярним інструментом, що використовується для надання огляду основних показників контролю якості необроблених даних секвенування наступного покоління. Методичні вказівки містять детальні інструкції до практикумів, завдяки чому студенти мають здобути навички роботи з програмою FastQC.

УДК 579.6+ 578

ЗМІСТ

ВСТУП

За останні кілька років з'явилося кілька платформ секвенування високої пропускної здатності (HTS) (або секвенування наступного покоління, NGS), які базуються на різних реалізаціях паралельного секвенування. Концепцію паралельного секвенування можна характеризувати як секвенування щільного масиву ознак ДНК ітераційними циклами ферментативних маніпуляцій та збору даних на основі зображень [Mitra et al., 1999; Behjati et al., 2013; Nilesh et al., 2021; Slatko et al., 2018] . Комерційні продукти, що базуються на цій технології секвенування, включають Roche's 454, Illumina's Genome Analyzer, ABI's SOLiD, Heliscope from Helicos та інші.

Поява платформ HTS відкрило багато можливостей для вивчення геномних варіацій [Campbell et al., 2008]. Секвенування цілих геномів має потенціал для відповіді на більшість ключових питань, пов'язаних з еволюцією, глобальною передачею та локальною адаптацією в бактеріях.

Розвиток сучасних інструментів біоінформатики допомогло подальшому розумінню еволюції та локальної та глобальної передачі декількох патогенних бактерій.

Однак, кожного разу після секвенування необхідно перевіряти якість отриманих послідовностей. Найбільш популярним інструментом є програма FastQC написана Саймоном Ендрюсом з Babraham Bioinformatics, яка використовується для надання огляду основних показників контролю якості необроблених даних секвенування наступного покоління.

Результатом FastQC після аналізу FASTQ-файлу читання послідовностей є html-файл, який можна переглянути в браузері. Звіт містить один розділ результатів кожного модуля FastQC. На додаток до графічних даних або списків, що надаються кожним модулем, призначається прапор "Пройдено", "Попередження" або "Помилка".

Мета методичних вказівок – ознайомити студентів з аналізом якості послідовності при використанні програми FastQC. Підготовка по цьому розділу дає можливість отримати теоретичну базу і практичні навички

використання специфічних серверів для обробки даних секвенування та навчить індивідуальній роботі з сервером Galaxy.

РОЗДІЛ 1. Технології NGS

1.1. Секвенування наступного покоління (NGS)

Секвенування нуклеїнових кислот - метод визначення нуклеотидної послідовності, що дозволяє отримати опис первинної структури лінійної макромолекули у вигляді послідовності мономерів (нуклеотидів).

Перші підходи для проведення секвенування були розроблені Едманом, потім Максамом, Гилбертом [Maxam et al., 1997] і Сенгером [Sanger et al., 1997]. Найбільшого поширення набув метод Сенгера, який відноситься до секвенування першого покоління і вважається «золотим стандартом», оскільки дозволяє визначати нуклеотидну послідовність досліджуваної ДНК з високою точністю [Voelkerding et al., 2009; Moorthie et al., 2011; Zhang et al., 2011].

Секвенуючи платформи, що використовують метод Сенгера, здійснюють читання нуклеотидної послідовності з високою точністю, але невеликою продуктивністю. В основі методу лежить терминування синтезу ланцюга ДНК ДНК-полімеразою за допомогою дідезоксірібонуклеозід трифосфату (рис. 1). Отримані фрагменти ДНК розділяють електрофорезом в поліакриламідному гелі за величиною фрагмента [Sanger et al., 1997].

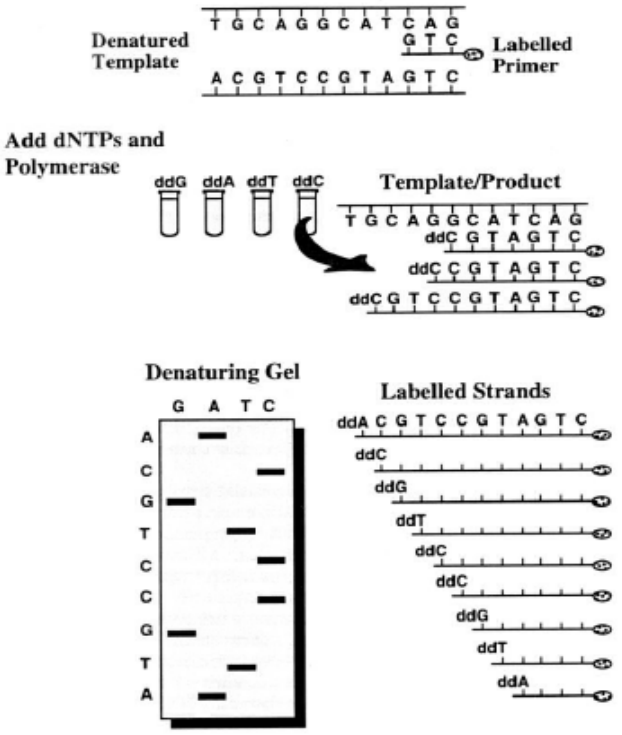


Рис. 1. **Технологія секвенування за Сенгером**

До теперішнього часу розроблені автоматичні капілярні секвенатори використовують метод Сенгера. Найбільшою популярністю користуються прилади виробника «Applied Biosystems». Їх останні моделі з 48 і 96 капілярами дозволяють за 2 год прочитати до 1100 нуклеотидів кожного зразка, що сумарно становить близько 5 Мб/день (www.appliedbiosystems.com.). Використання флуоресцентних міток, кожна з яких відповідає одному з чотирьох нуклеотидів дозволяє проводити детекцію результатів реакції за допомогою лазера [Pettersson et al., 2009; Voelkerding et al., 2009; Pareek et al., 2011; Moorthie et al., 2011].

Новий, розроблений під час проведення проекту «Геном Людини» спосіб визначення більш довгих послідовностей ДНК, названий «shotgun-sequencing» (метод «дробовика»), послужив основою для масивного паралельного секвенування, використовуваного в NGS [Zhang et al., 2011]. При використанні даного методу геномної ДНК ензиматичними або хімічним способом гідролізується на короткі фрагменти, клонують і використовуються в подальшому для визначення нуклеотидної послідовності методом Сенгера.

Повну послідовність досліджуваного фрагмента ДНК визначають шляхом вирівнювання і об'єднання отриманих нуклеотидних послідовностей за рахунок їх часткового перекривання.

Вперше розроблене в 1990 році, капілярне секвенування паралельно розподіляло традиційне секвенування Сенгера в пристрої, які могли виконувати до 384 реакцій одночасно [Ku et al., 2012]. Використання радіоізотопних міток замінювалося флуоресцентним детектуванням; гелі поліакриламідних плит замінювали полімерно-заповненими капілярами, а рентгенівську плівку залишали на користь лазерної флуоресцентної детекції.

Технології NGS дозволяють одночасно визначати нуклеотидні послідовності безлічі різних ниток ДНК при здійсненні процесів синтезу або лігування ДНК, що забезпечує читання мільярдів нуклеотидів в день [Barzon et al., 2011; Pettersson et al., 2009; Voelkerding et al., 2009; Pareek et al., 2011; Moorthie et al., 2011; Zhang et al., 2011; Caporaso et al., 2012].

Масивне паралельне секвенування, інакше зване секвенуванням наступного покоління «next-generation sequencing (NGS)», відноситься до методів, які з'явилися в останнє десятиліття.

Принципова відмінність технологій NGS від методу термінації зростаючого ланцюга, розробленого Фредеріком Сенгером [Sanger et al., 1997], полягає в можливості паралельного визначення нуклеотидних послідовностей безлічі різних ниток ДНК і читання мільярдів нуклеотидів в день [Pareek et al., 2011; Moorthie et al., 2011; Zhang et al., 2011].

Завдяки різноманітних поліпшень, численних модифікацій і збільшення продуктивності обчислювальної техніки було розроблено технології секвенування, які ґрунтуються на одноразовому «прочитанні» (розпаралелювання) кількох ділянок геному.

Технології секвенування нового покоління дозволяють прочитати одноразово кілька ділянок геному, що є головною відмінністю від попередніх технологій.

З появою платформ NGS відкриваються нові перспективи у вивченні повногеномних послідовностей великої кількості патогенів (наприклад, M. *tuberculosis*, *E. coli*, *H. pylori*, *S. aureus*, *V. cholerae*, *N. meningitides*, вірусів парентеральних гепатитів, вірусів групи герпесу тощо), в створенні інформаційних технологій та методів молекулярно-генетичної характеристики збудників інфекційних захворювань, що дозволяють вирішувати як фундаментальні, так і практичні завдання мікробіології, вірусології, епідеміології [Мокроусов, 2012; Соломатіна и др., 2012; Koser et al., 2012; Harris et al., 2013; Попов и др., 2009; Gardy et al., 2009; Eyre et al., 2012; Hasan et al., 2009; Grad et al., 2009].

Є досвід використання повногеномного секвенування в епідеміологічному розслідуванні спалахів холери, туберкульозу та ешеріхіоза (діареї з гемолітико-уремічний синдром), обумовленого ентерогеморрагічним штамом E. coli O104: H4 [Gardy et al., 2009, Hasan et al., 2009, Grad et al., 2009].

Основними точками програми методів NGS в галузі мікробіології, вірусології та епідеміології є:

• відкриття нових бактерій і вірусів з використанням метагеномногих підходів;

• вивчення мікробних спільнот різних біотопів тіла здорової людини і в стані хвороби;

• аналіз варіабельності геномів збудників інфекційних захворювань.

Процес секвенування на платформах NGS складається з декількох етапів. На першому етапі здійснюють процес підготовки бібліотеки ДНК, який включає фрагментованість ДНК ферментативно або за допомогою ультразвуку з подальшим приєднанням до отриманих фрагментів ДНК універсальних олігонуклеотидних адаптерів відомої послідовності і індексів за допомогою полімеразної ланцюгової реакції (ПЛР). Адаптери необхідні для подальшої ампліфікації фрагментів.

В даний час для високопродуктивного секвенування доступні кілька комерційних платформ: HiSeq, MiSeq і NextSeq 500 (Illumina), Roche 454 GS, Ion torrent (Thermo Fisher Scientific) і SOLiD (Applied Biosystems) [Cronin et al., 2011; Meldrum et al., 2011; Desai et al., 2012; Ku et al., 2012] (табл. 1).

Таблиця 1

Порівняльна характеристика відомих секвенаторів

Платформа	**Illumina MiSeq**	**Illumina GAIIx**	**Illumina HiSeq 2000**	**Ion Torrent PGM**	**PacBio Rs**	**454 GS FLX**
Принцип	SBS	SBS	SBS	Іоний полупров.	Real-time	Пиросекв.
Кошта прибора	$ 128k	$256k	$654k	$80k	$695k	$500k
Час работи за цикл	27ч	10д	11д	2ч	2ч	23ч
Довжина читання	до 150 нукл	до 150 нукл	до 150 нукл	~200 нукл	~2500 нукл	700 нукл
Кількість нуклеотидів за цикл	1.5-2 Гн	30 Гн	600 Гн	20-1000 Мн	100 Мн	700 Мн

Продовження таблиці 1

Кошта 1Г нукл.	$502	$148	$41	$1000	$2000	$7000
Частка помилок	0.8%	0.76%	0.26%	1.71%	12.86%	0.49%
Головний тип помилок	Заміна	Заміна	Заміна	Вставка/ видалення	Вставка/ видалення	Вида-ленн я/встав-ка

Другий етап полягає в проведенні ампліфікації кожного фрагмента ДНК методом ПЛР. Фрагмент ДНК за допомогою послідовності адаптера гибрідизуються з одним або двома праймерами, іммобілізованими на твердій поверхні (Мікрокульки або скляний чіп) і беруть участь в ПЛР. Через чіп (проточна комірка) пропускається реакційна суміш, яка містить набір ферментів для секвенування. Далі відбувається автоматичне покрокове зчитування кожного типу нуклеотиду і детекція результату [Pettersson et al., 2009; Moorthie et al., 2011].

Фінальними етапами роботи є безпосередньо секвенування і аналіз отриманих даних.

Секвенування проводиться шляхом синтезу нових фрагментів ДНК на одноланцюгових ДНК-бібліотеках, які виконують роль матриці. Нуклеотиди вбудовуються в новий ланцюг в певному порядку, відповідному матричному ланцюгу, який записується в цифровому вигляді. Після включення в ланцюг кожного наступного нуклеотиду прилад реєструє сигнал. Різні платформи для NGS використовують різні механізми детекції вбудованих в новий ланцюг нуклеотидів: детекцію флуоресцентного сигналу після включення в ланцюг комплементарної матриці нуклеотиду (Illumina, SOLiD Applied Biosystems), зміна pH розчину в мікрореакторі, пов'язане з виділенням в середу іонів водню в ході синтезу ДНК ферментом(Ion torrent Thermo Fisher Scientific), або реєстрацію світлового сигналу після виділення пірофосфату, що активує каскад хімічних реакцій (Roche) [Quail et al., 2012].

1.2. Переваги та недоліки NGS у порівнянні з методом Сенгера

В цілому секвенатори на основі нових технологій стали значно ефективніше і набагато дешевше своїх попередників [Nielsen, 2010] .

В даний час продуктивність деяких секвенаторов перевершує сотні мільярдів нуклеотидів за робочий цикл, що дає можливість таким приладам лише за кілька днів отримати геном людини.

У принципі, концепції, що стоять як за технологіями секвенування наступного покоління (NGS), так і за секвенуванням методом Сенгера подібні. В секвенуванні як NGS, так і Sanger, ДНК-полімераза додає флуоресцентні нуклеотиди один за одним на зростаючу нитку ДНК. Кожен вбудований нуклеотид ідентифікують за його флуоресцентною міткою [Taub et al., 2010] .

На ринку існує досить велика кількість секвенаторів, заснованих на різних принципах визначення нуклеотидів. Тому всі вони розрізняються в

швидкості і вартості секвенування, вартості самого приладу, супутніх типах помилок і їх кількості [Quail et al., 2012].

Критична різниця між секвенуванням Sanger і NGS полягає в об'ємі секвенованого матеріалу Коли методом Сенгера секвенується лише один фрагмент ДНК одночасно, більшість NGS платформ здатне одночасно секвенувати мільйони фрагментів ДНК за один цикл. NGS також пропонує більшу здатність виявляти нові або рідкісні варіанти з глибоким секвенуванням (табл. 2).

У всіх платформах останнього покоління для зчитування нуклеотидних послідовностей (в порівнянні з попередніми аналогами) за більш високу швидкість і низьку вартість обробки даних доводиться розплачуватися більш частими помилками секвенування [Hoff, 2009]. Але крім звичайних випадкових розбіжностей існують ще й систематичні джерела помилок. У деяких геномних позиціях в багатьох читаннях трапляються скупчення розбіжностей з референсними нуклеотидами. Відомо, що помилки відбуваються ближче до кінців читань [2008, Taub et al., 2010, Nakamura et al., 2011] і залежать від оточуючих її мотивів послідовності. Наприклад, помилки схильні до появи перед «GG» або після деякої кількості «GGC» [Nakamura et al., 2011]. Було виявлено [Meacham et al., 2011], що є геномні позиції, в яких помилки трапляються набагато частіше, ніж це може бути пояснено описаними ефектами. Такі помилки мають систематичний характер появи. Було відмічено, що в місцях скупчення систематичних помилок, як правило, помилки з'являються тільки з одного боку секвенування (пряма або зворотна).

Після секвенування отримані дані можуть бути оброблені або за допомогою біоінформатика, або - спеціального програмного забезпечення, встановленого на приладі або сервері. Дані проходять кілька етапів обробки: виключення рідов з низькою якістю читання, вирівнювання даних щодо референсної послідовності або складання послідовності de novo.

Таблиця 2

Порівняння секвенування за Sanger та NGS

	Sanger Sequencing	NGS
Переваги	● Швидке, економічне секвенування для низького числа цілей (1–20 цілей ● Знайомий робочий процес	● Вища глибина секвенування дозволяє підвищити чутливість (до 1%) ● Вища здатність знаходити нові варіанти ● Вища здатність ідентифікувати мутації ● Більше даних, отриманих з однаковою кількістю вхідної ДНК ● Вища пропускна спроможність
Недоліки	● Низька чутливість (межа виявлення ● ~ 15–20%) ● Низька потужність виявлення ● Не є економічно ефективним для великої кількості цілей (> 20 цілей) ● Низька масштабованість завдяки збільшенню вимог до введення зразків	● Менш рентабельний для визначення низької кількості цілей (1–20 цілей) ● Тривалий час для визначення низької кількості цілей (1–20 цілей)

1.3. Предпроцесингові заходи

Мета фільтра предпроцесингової обробки - виправити або усунути хибні читання перед початком процесу складання. Ці помилки викликані платформами послідовності і, отже, вони різняться між платформами. Виявлення і виправлення цих помилок на ранньому етапі полегшить процес складання і запобіжить неправильні збірки на більш пізніх етапах. Алгоритми корекції помилок варіюються від простих процесів обрізки з використанням базових оцінок якості до складних підходів до корекції помилок на основі частоти помилкових зчитувань в збираємому наборі [Brown et al., 2017]. Всі алгоритми виправлення помилок засновані на тій же загальній концепції, що читання з помилками неефиктивно і випадково, тому їх можна виявити, підрахувавши читання в пулі збірок. Низькочастотні зчитування є кандидатами для алгоритмів корекції помилок і можуть бути виправлені з високочастотними зчитуванями. Однак на цю ідею впливають проблеми високочастотних геномних повторів і неоднорідної вибірки генома, які призводять до неоднозначних результатів, одержуваних з декількох рівних варіантів корекції.

Перший з модулів програми FastQC [Andrews, 2009] показує огляд діапазону значень якості по всіх базах в кожній позиції в файлі fastq. Для кожної позиції рисується графік типу BoxWhisker де:

• середня якість (синя лінія)

• медіана - таке значення якості, що половина рідов по даній позиції мають якість не вище зазначеного (центральні червоні лінії)

• інтерквартільний розмах - різниця між верхньою і нижньою квартилями - такий діапазон значень якості, що 25% читань мають значення якості по даній позиції не вище нижньої межі, а 75% читань - не вище верхньої (жовтий прямокутник).

• розмах між 10-й і 90-й процентиль - 10% читань мають значення якості не вище нижньої межі, 90% - не вище верхньої.

Ось Y на графіку показує показники якості. Залежно від цього поле графіка розділене на смуги різних кольорів (зелена, жовта, червона), потрапляння вищезазначених елементів які дозволяють судити про якість читань (Ця інформація базується на очевидному факті: чим вище значення якості, тим краще) (рис. 2).

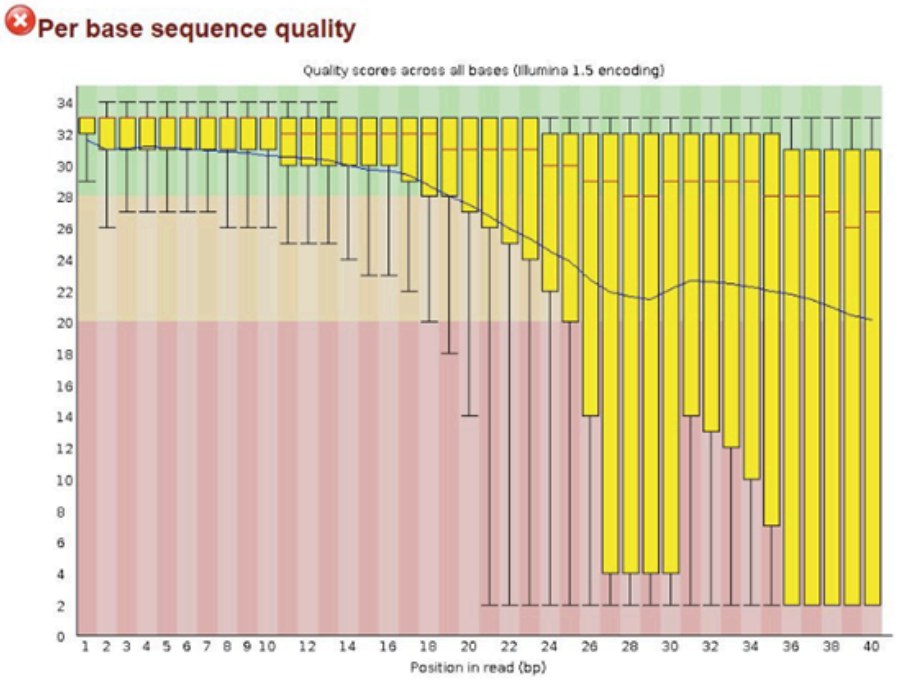


Рис. 2. **Приклад плохої якості секвенування**

Per sequence quality scores - графік загального числа читань у порівнянні із середньою оцінкою якості по всій довжині цього читання (рис. 3). Звіт за показниками якості для кожної послідовності дозволяє побачити, чи мають підмножина ваших послідовностей універсально низькі значення якості. Розподіл середньої якості читання має бути досить вузьким в верхньому діапазоні графіка [Andrews, 2009].

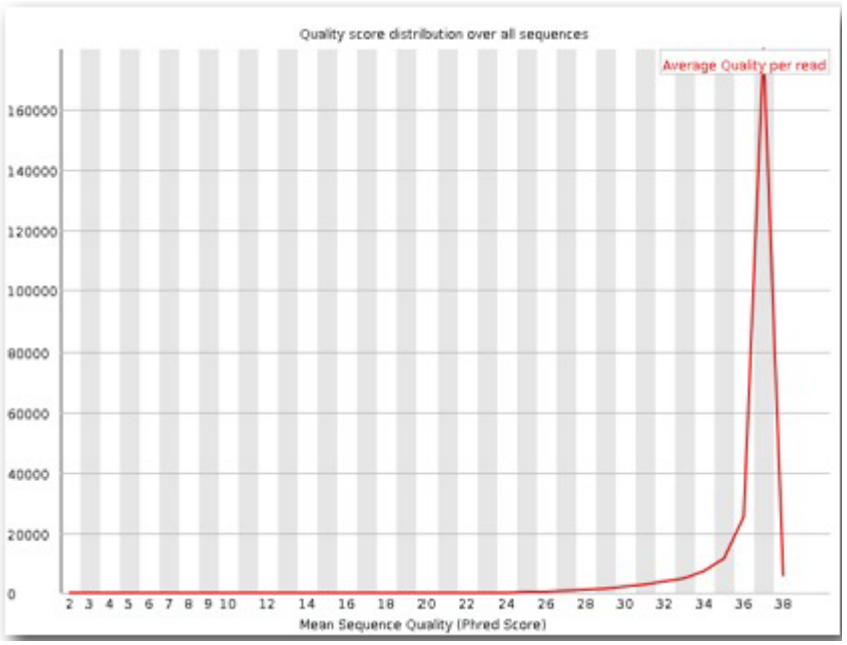
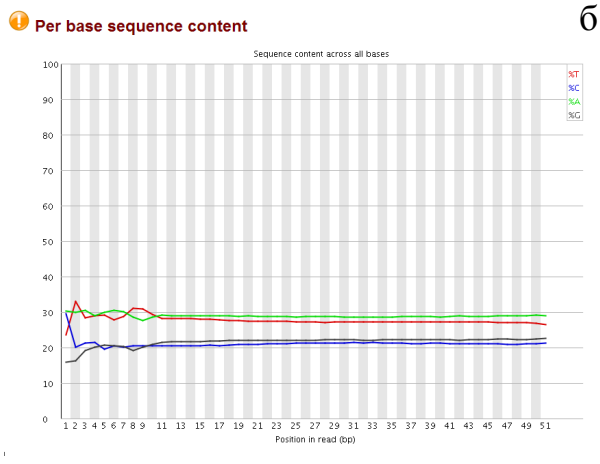


Рис. 3. **Приклад гарного розподілу якості читання**

Per Base Sequence Content на цьому графіку представлений відсоток баз, викликаних для кожного з чотирьох нуклеотидів в кожній позиції, у всіх читаннях в файлі (рис. 4). В випадкової бібліотеці ви очікуєте, що між різними основами послідовності не буде практично ніякої різниці, тому лінії на цьому графіку повинні проходити паралельно один одному. Відносна кількість кожної основи повинна відображати загальна кількість цих основ у вашому геномі, але в будь-якому випадку вони не повинні бути сильно розбалансовані один від одного, тобто повинна дотримуватися приблизна пропорція % A =% T і % G =% C на всьому протязі довжини зчитування. Однак, при використанні більшості протоколів отримання бібліотеки

RNA-Seq спостерігається чіткий нерівномірний розподіл основ для перших 10-15 нуклеотидів; це нормально і очікується в залежності від типу використовуваного набору бібліотек. Якщо ви бачите сильні відхилення, які змінюються в різних основах, то це зазвичай вказує на надмірно представлену послідовність, яка забруднює вашу бібліотеку. Зсув, який однаковий для всіх основ, вказує або на те, що вихідна бібліотека була зміщена по послідовності, або, під час секвенування бібліотеки виникла систематична проблема [Andrews, 2009].

а



б

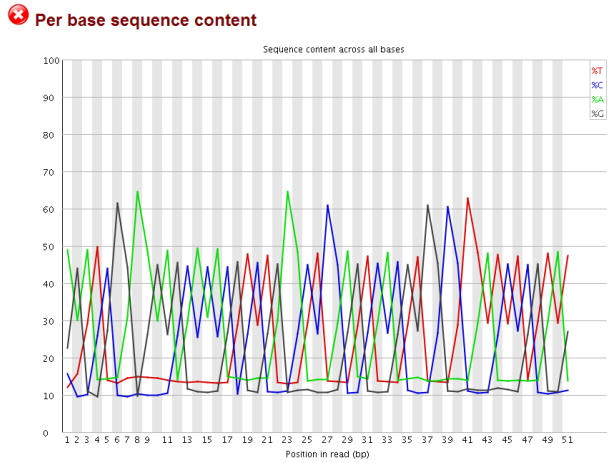


Рис. 4. **Приклад якісного (а) та не якісного (б) розподілу нуклеотидів у читанні**

Per Base GC Content відображає теоретичний розподіл, який передбачає однорідний зміст GC для всіх операцій читання (рис. 5). Для повногеномного секвенування очікується, що зміст GC всіх операцій читання має сформувати нормальний розподіл з піком кривої при середньому вмісті GC в організмі. Якщо спостерігаємий розподіл відхиляється занадто далеко від теоретичного, то це може вказувати на надмірно представлену послідовність, яка забруднює вашу бібліотеку. Зсув, який однаково для всіх основ, вказує або на те, що вихідна бібліотека була зміщена по послідовності, або на те, що під час секвенування бібліотеки виникла систематична проблема [Andrews, 2009].

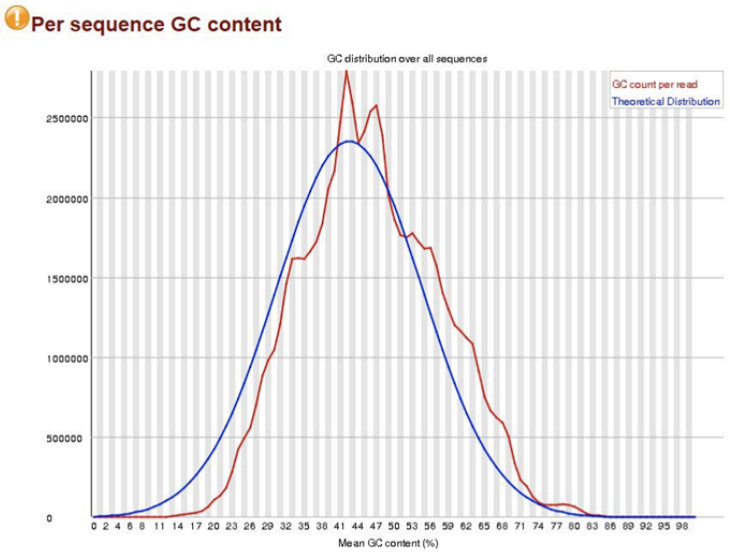


Рис. 5. **Приклад загрязнення послідовності**

Sequence Duplication Levels показує відсоток читань певної послідовності в файлі, які присутні в файлі певну кількість разів (це синя лінія) (рис. 6). В даному випадку очікується, що майже 100% ваших читань будуть унікальними (з'являються тільки 1 раз в даних послідовності). Це вказує на дуже різноманітну бібліотеку, яка не була надмірно впорядкована. Якщо вихідні дані секвенування дуже глибокі (наприклад,> 100X розмір вашого генома), ви почнете бачити деякі дублювання послідовності; це неминуче, оскільки в теорії існує тільки кінцеве число повністю унікальних зчитувань послідовностей, які можна отримати з будь-якого заданого вхідного зразка ДНК. При секвенуванні РНК будуть присутні деякі дуже рясні транскрипти. Очікується, що повторні читання будуть спостерігатися для транскриптів з високим вмістом [Andrews, 2009].

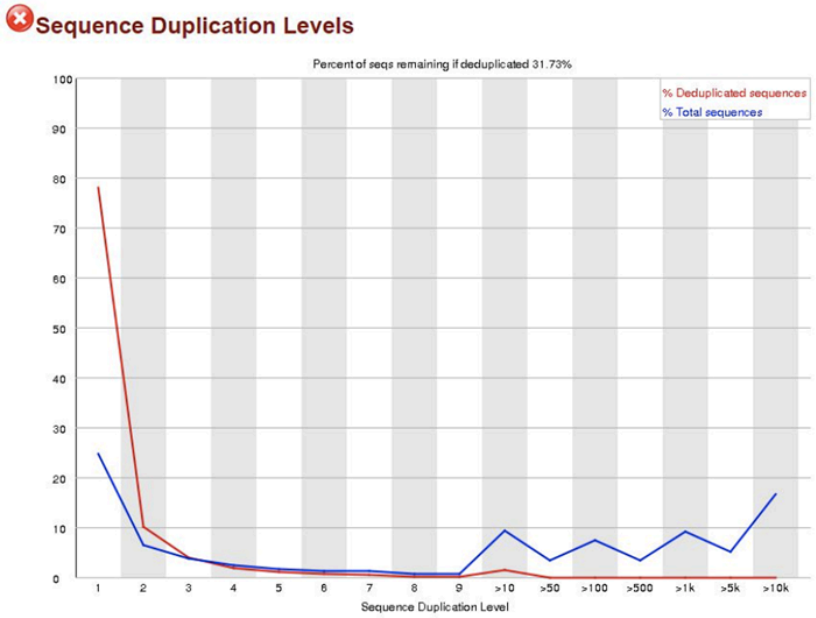


Рис. 6. **Приклад присутності багато повторюючихся послідовностей**

Overrepresented Sequences показує список послідовностей, які відображаються в файлі більше, ніж очікується (рис. 7). Вважається тільки перші 50 bp. Послідовність вважається перепредставленою, якщо вона становить ≥0,1% від загального числа читань. Кожну надмірно представлену послідовність порівнюють зі списком поширених забруднюючих речовин, щоб спробувати її ідентифікувати. Виявлення того, що одна послідовність дуже перепредставлена в наборі, означає або її високу біологічну значимість, або вказує на те, що бібліотека забруднена, або не настільки різноманітна, як ви очікували. В даних DNA-Seq не повинно бути присутнім жодної послідовності з досить високою частотою, щоб її можна було вважає перепредставленою, хоча нерідко можна бачити невеликий відсоток, який вказує на читання адаптера [Andrews, 2009].

Рис. 7. **Приклад виявлених перепредставлених послідовностей**

РОЗДІЛ 2. Робота в FastQC

Відкриємо сторінку сервера Galaxy (https://usegalaxy.org.au/) (рис 8.)

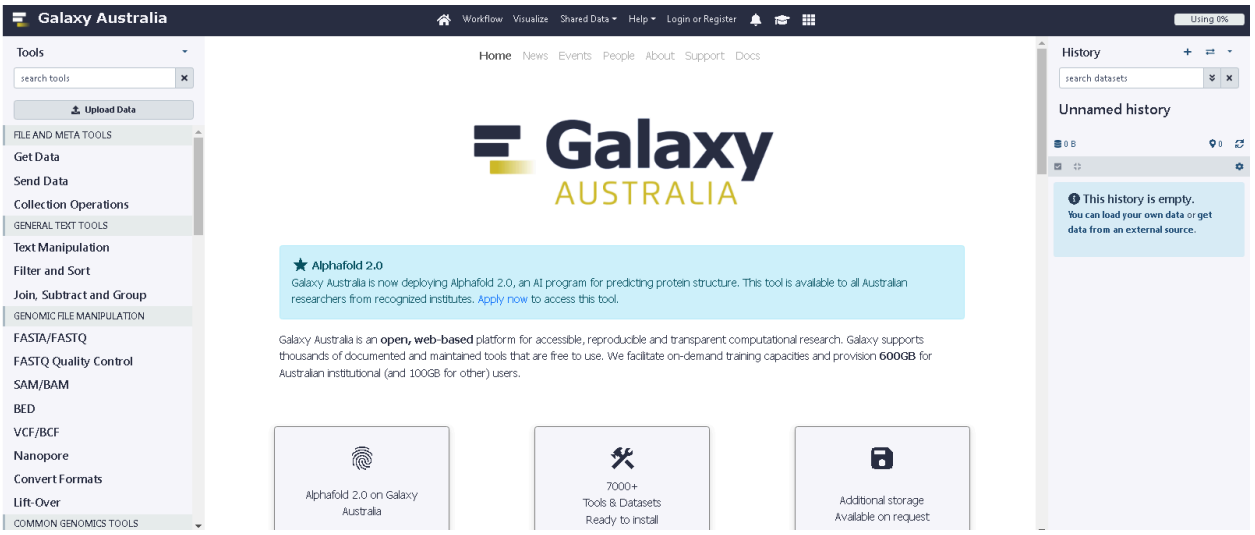


Рис. 8 **Головна сторінка серверу Galaxy** (https://usegalaxy.org.au/)

Слід зауважити, що ми можете працювати з інструментом **FastQC** на бідь якому сервері Galaxy, що містить блок метагеномного або біоінформатичного аналізу.

Потім необхідно завантажити ваші данні. Якщо вони зберігаються на локальному диску, то відкрийте диспетчер завантажень **Upload Data** (1), потім **Choose local files** (2) і **Start** (3). Після завантаження натисніть **Stop** (4). Зовнішній вигляд активної панелі наведено на рисунку 9.

Якщо ваші дані зберігаються на сервері, то по-перше скопіюйте розташування посилання, потім відкрийте диспетчер завантажень **Upload Data**, оберіть **Paste/Fetch Data** та вставте посилання в текстове поле і натисніть **Start.** Після завантаження натисніть **Stop**

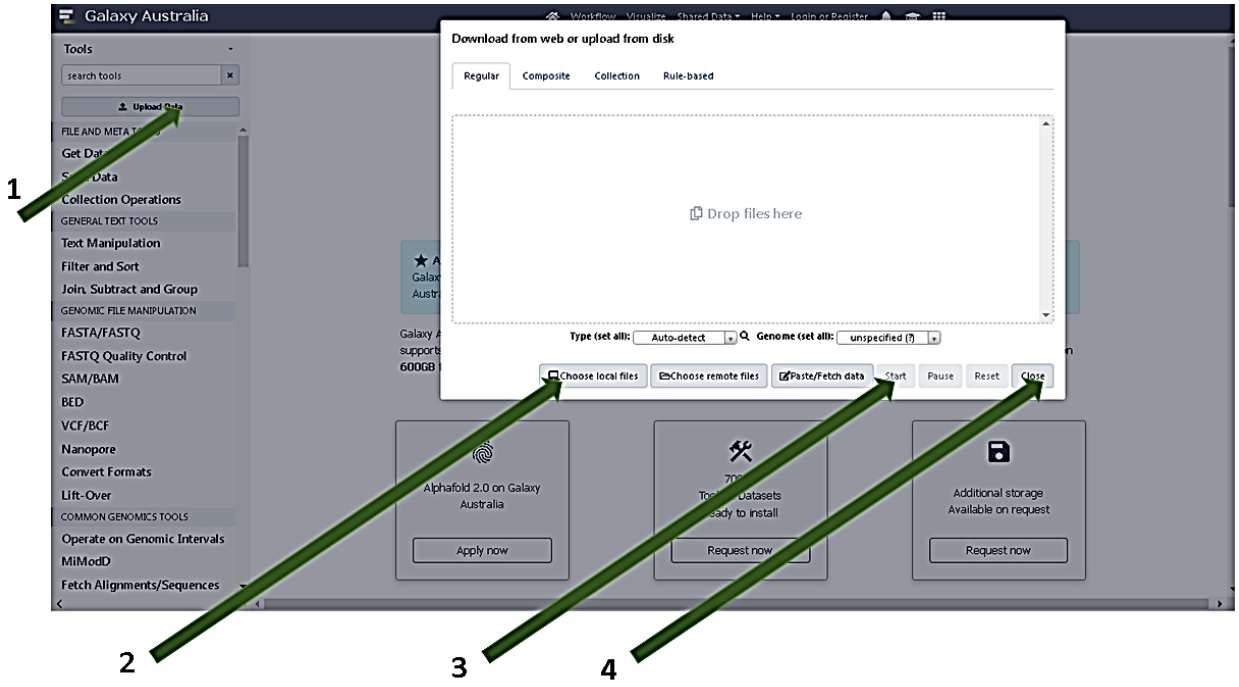


Рис. 9. **Панель для завантаження файлів з локального диску на сервер Galaxy**

Після того, як ваші дані завантажені ви можете їх переглянути (рис. 10) для чого необхідно натиснути на символ перегляду (значок очі) на панелі поряд з позначенням завантаженого файлу (рис. 10)

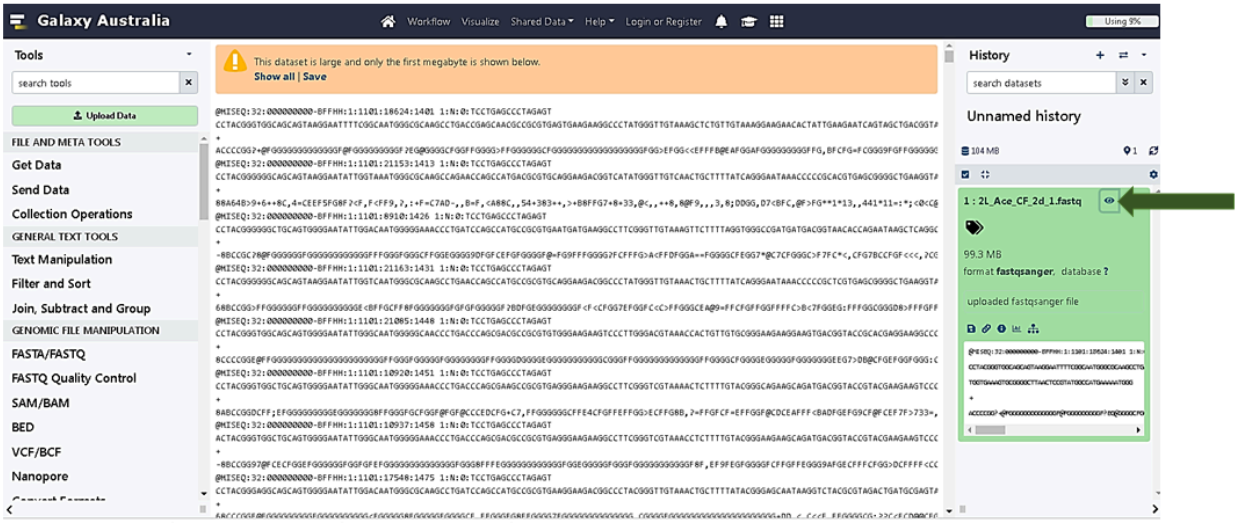


Рис. 10. **Розгорнутий вигляд завантаженого файлу формату FASTQ**

Кожне читання файлу формату FASTQ, яке представляє фрагмент бібліотеки, кодується 4 рядками:

Опис

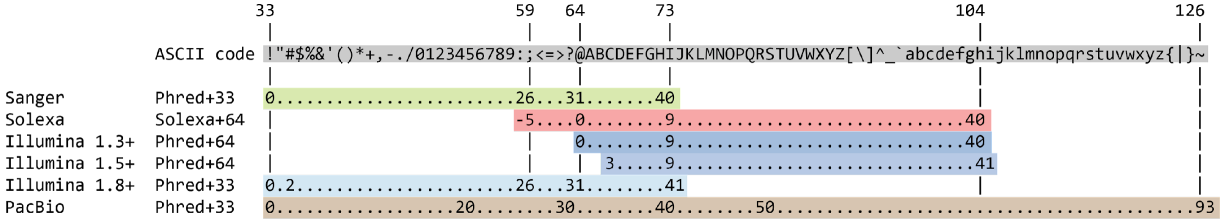
1 Завжди починається з @, за яким слідує інформація про прочитане

2 Фактична нуклеїнова послідовність

3 Завжди починається з +і іноді містить ту саму інформацію в рядку 1

4 Має рядок символів, який представляє показники якості, пов'язані з кожною основою нуклеїнової послідовності; повинна мати таку ж кількість символів, що й рядок 2

Показник якості для кожної послідовності є рядком символів, по одному для кожної основи нуклеїнової послідовності, використовувану для характеристики ймовірності помилкової ідентифікації кожної основи. Оцінка кодується за допомогою таблиці символів ASCII



Таким чином, з кожним нуклеотидом пов'язаний символ ASCII, що представляє його показник якості Phred, ймовірність неправильного базового виклику:

Показник якості Phred	Можливість неправильного базового виклику	Базова точність виклику
10	1 из 10	90%
20	1 из 100	99%
30	1 из 1000	99,9%
40	1 из 10 000	99,99%
50	1 из 100 000	99,999%
60	1 из 1 000 000	99,9999%

Запуск FASTQC

У полі пошуку панелі інструментів знайдіть "FastQC". Інтерфейс інструмента відобразиться на центральній панелі Galaxy (рис. 11).

• У полі **Raw read data from your current history** має з'явитися ваша послідовність (1)

• Якщо у вас є **список адаптерів, його необхідно додати в історію, а потім у рядок Adapter** list панелі FastQC. Якщо даних немає, залиште поле пустим (2)

• Натисніть **Execute** (3)

• На панелі «**History**» клацніть піктограму «Оновити», щоб побачити, чи завершено аналіз.

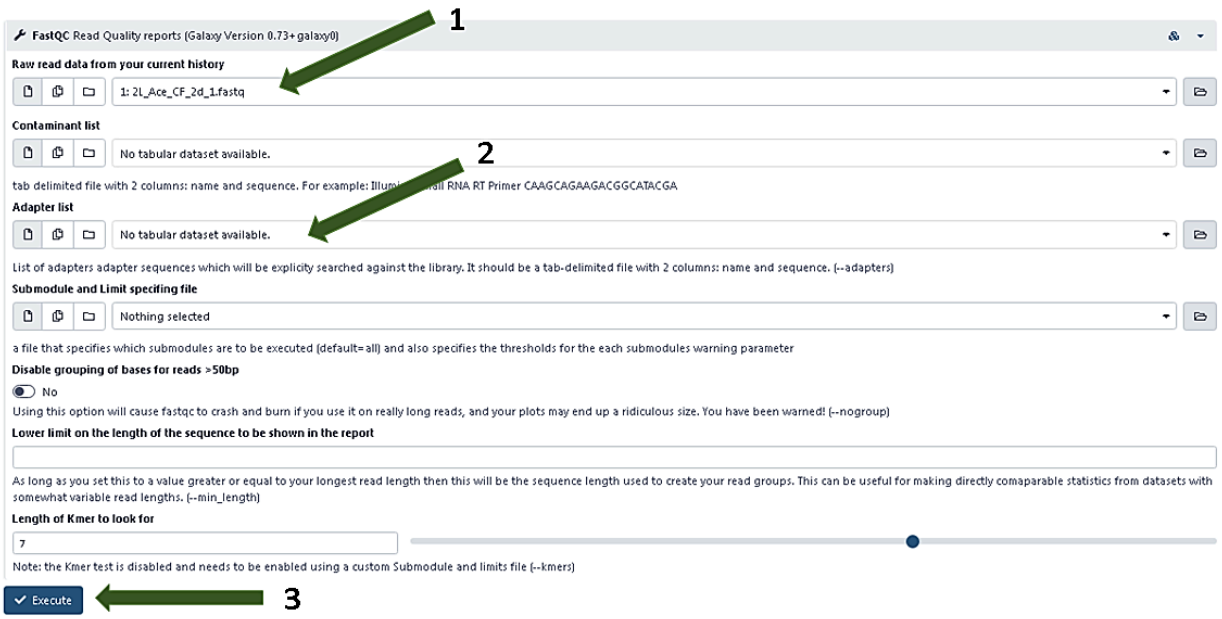


Рис. 11. **Активна панель FastQC**

Після завершення перевірте висновок, званий **FastQC on data1:Webpage** (клікніть значок очі). У верхній частині сторінки є свідка і кілька графіків.

Дивитися на:

Basic Statistics

Sequence length: буде важливо при встановленні максимального значення розміру k-mer для збирання.

Encoding: Тип якісного кодування важливий для якісного обрізання програмного забезпечення.

% GC: організми з високим GC не схильні добре збиратися та можуть мати нерівномірний розподіл покриття читання.

Total sequences: Загальна кількість прочитань: дає уявлення про охоплення.

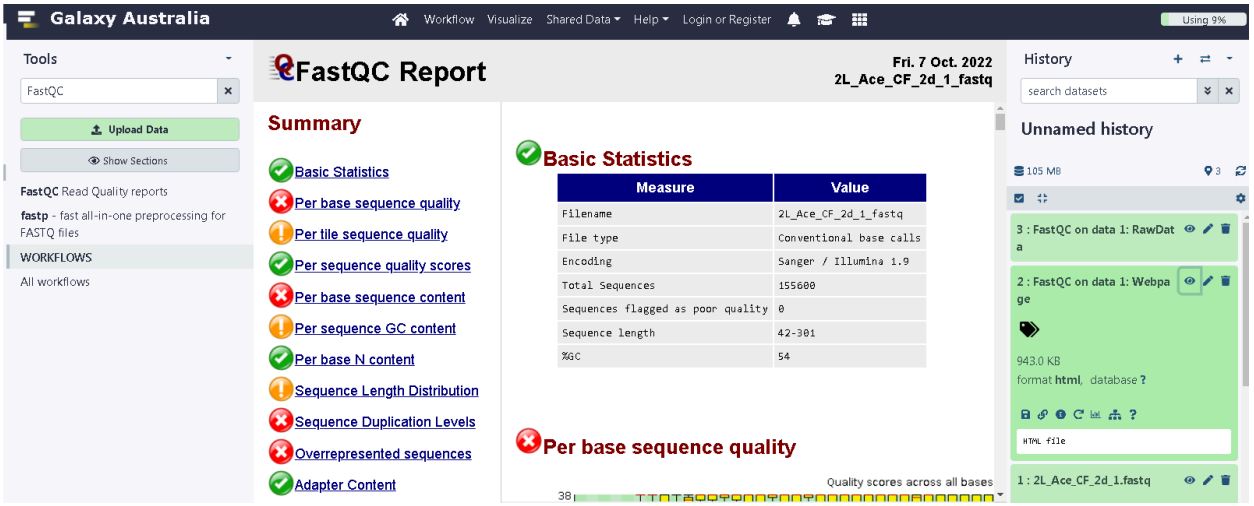


Рис. 12. **Розділ Basic Statistics для послідовності, що аналізуємо в FastQC**

Per tile sequence quality. Цей графік з'явиться у результатах аналізу лише у тому випадку, якщо ви використовуєте бібліотеку Illumina, у якій збережено вихідні ідентифікатори послідовностей. У них закодовано комірку проточної кювети, з якого було отримано кожне зчитування. Графік дозволяє переглянути показники якості кожної комірки на всіх ваших базах, щоб побачити, чи втрата якості була пов'язана тільки з однією частиною проточної кювети. На графіці показано відхилення від середньої якості кожної комірки. Кольори знаходяться в діапазоні від холодного до гарячого, при цьому холодні кольори відповідають позиціям, де якість була на рівні або вище середнього для цієї бази в пробігу, а тепліші кольори вказують на те, що комірка має гірші якості, ніж інші комірки для цієї бази.

Цей модуль видасть попередження, якщо будь-яка плитка показує середній бал Phred більш ніж на 2 менше від середнього значення для цієї бази по всіх комірках.

Цей модуль видасть попередження, якщо будь-яка комірка показує середній бал Phred більш ніж на 5 менше ніж середнє значення для цієї бази по всіх комірках.

Як ми бачимо, для нашої послідовності, майже по всіх комірках показники якості були досить високими (рис. 13)

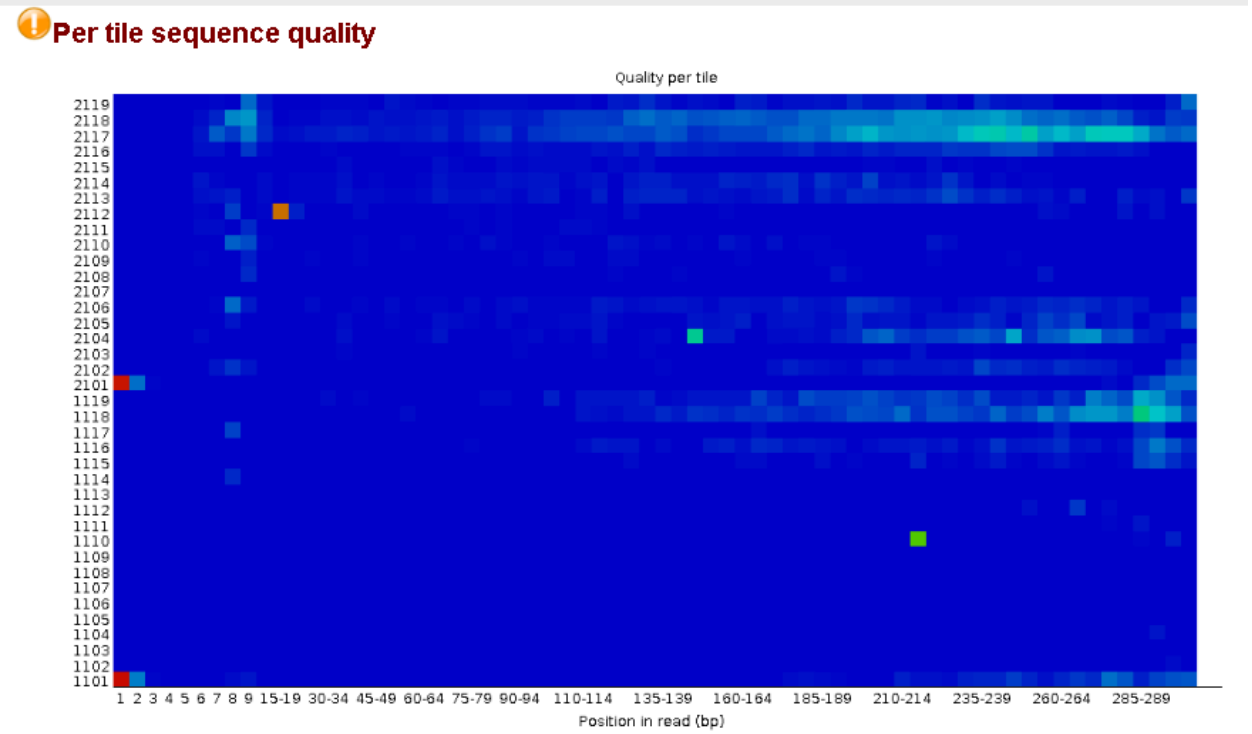


Рис. 13. Показники якості кожної комірки

Per base sequence quality: Провали якості спочатку можуть вказувати на технічні проблеми з процесом секвенування/запуском машини.

Попередження буде видано, якщо нижній квартиль для будь-якої основи менше 10 або якщо медіана для будь-якої основи менше 25.

Цей модуль викликає відмову, якщо нижній квартиль для будь-якої основи менше 5 або якщо медіана для будь-якої основи менше 20.

Для нашої послідовності риди високої якості за своєю довжиною становлять приблизно 150 п.н. (рис. 14).

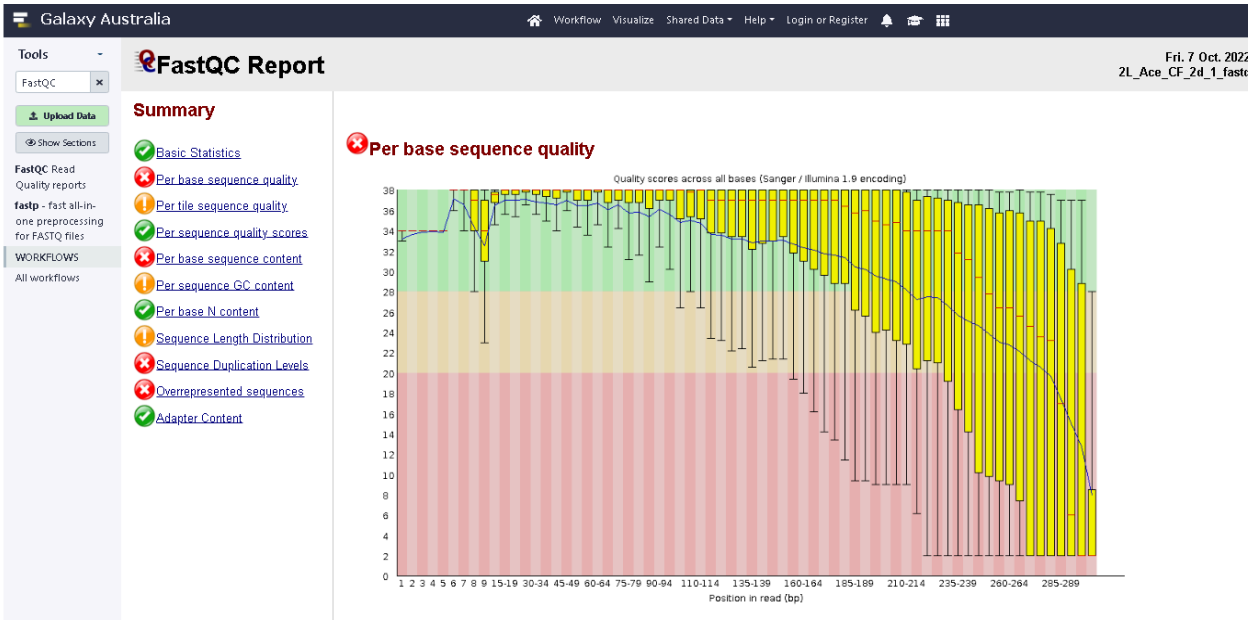


Рис. 14 **Графік розподілу якості секвенування**

Per sequence quality scores. Розподіл середньої якості читання має бути вузьким піком у верхньому діапазоні графіка.

Попередження видається, якщо середня якість, що найбільш часто спостерігається, нижче 27, що відповідає частоті помилок 0,2%.

Помилка виникає, якщо середня якість, що найчастіше спостерігається, нижче 20 — це відповідає частоті помилок 1%.

Для нашої послідовності показники якості досить гарні (рис. 15).

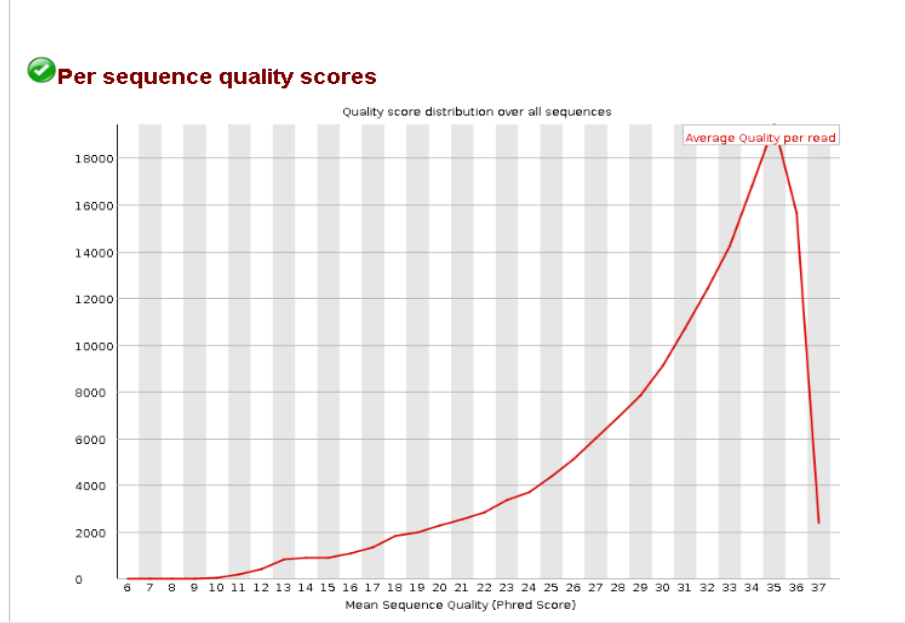


Рис. 15 **Приклад розподілу якості секвенування**

Per base sequence content. Як ми бачимо, на початку послідовностей зміст послідовності на підставу не дуже хороший, і відсотки не рівні, як очікувалося даних амплікону 16S. Варто зазначити, деякі типи бібліотек завжди будуть створювати упереджену композицію послідовності, зазвичай на початку читання. Бібліотеки, отримані шляхом праймування з використанням випадкових гексамерів (включаючи майже всі RNA-Seq) та ті, що були фрагментовані з використанням транспозаз, будуть містити внутрішнє зміщення у положеннях у перші 10-12 баз. Це зміщення не пов'язане з конкретною послідовністю, а натомість забезпечує збагачення ряду різних K-мірів на 5'-кінці ланцюга. Хоча це справді технічна помилка, її не можна виправити обрізанням, і в більшості випадків вона не має негативного впливу на якість.

Цей модуль видає попередження, якщо різниця між A та T або G і C перевищує 10% у будь-якому положенні.

Цей модуль вийде з ладу, якщо різниця між A та T або G і C перевищує 20% у будь-якій позиції.

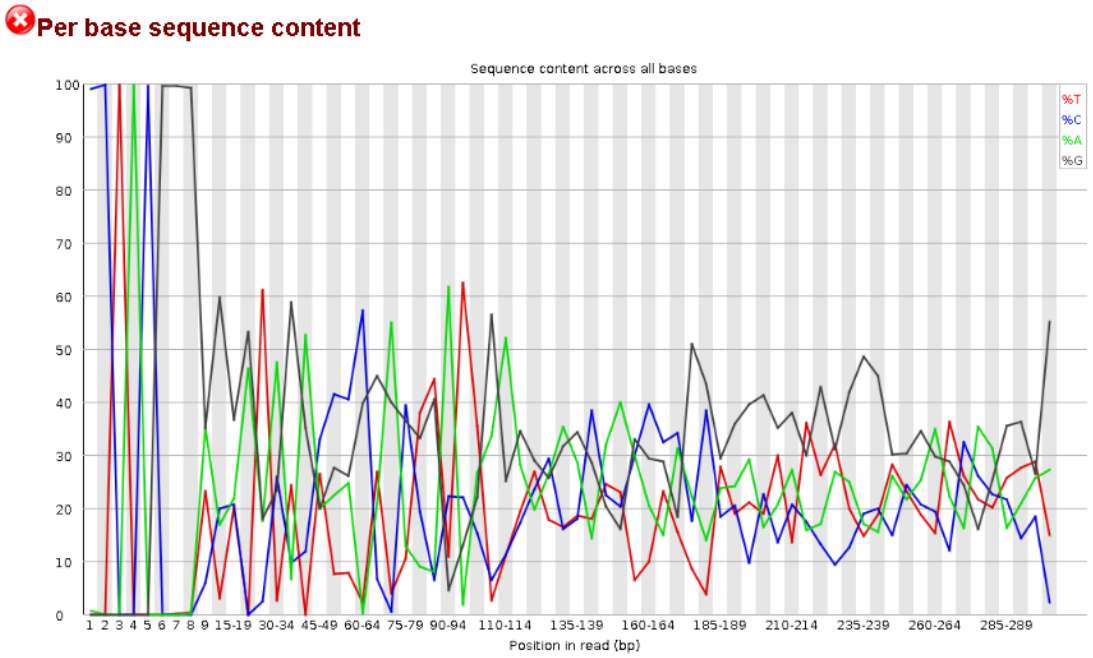


Рис. 16. **Графік розподілу нуклеотидів у читанні**

Per sequence GC content. Цей модуль вимірює вміст GC по всій довжині кожної послідовності у файлі та порівнює його зі змодельованим нормальним розподілом вмісту GC. У звичайній випадковій бібліотеці ви очікуєте побачити приблизно нормальний розподіл змісту GC, де центральний пік відповідає загальному змісту GC основного геному. Оскільки ми не знаємо змісту GC в геномі, модальний вміст GC розраховується на основі даних, що спостерігаються, і використовується для побудови еталонного розподілу.

Попередження видається, якщо сума відхилень від нормального розподілу становить понад 15% зчитувань.

Цей модуль вкаже на збій, якщо сума відхилень від нормального розподілу становить понад 30% зчитувань..

Як ми бачимо на рисунку 17 формування додаткових піків (червона лінія) може свідчити про сторонню контамінацію дослідженого зразка.

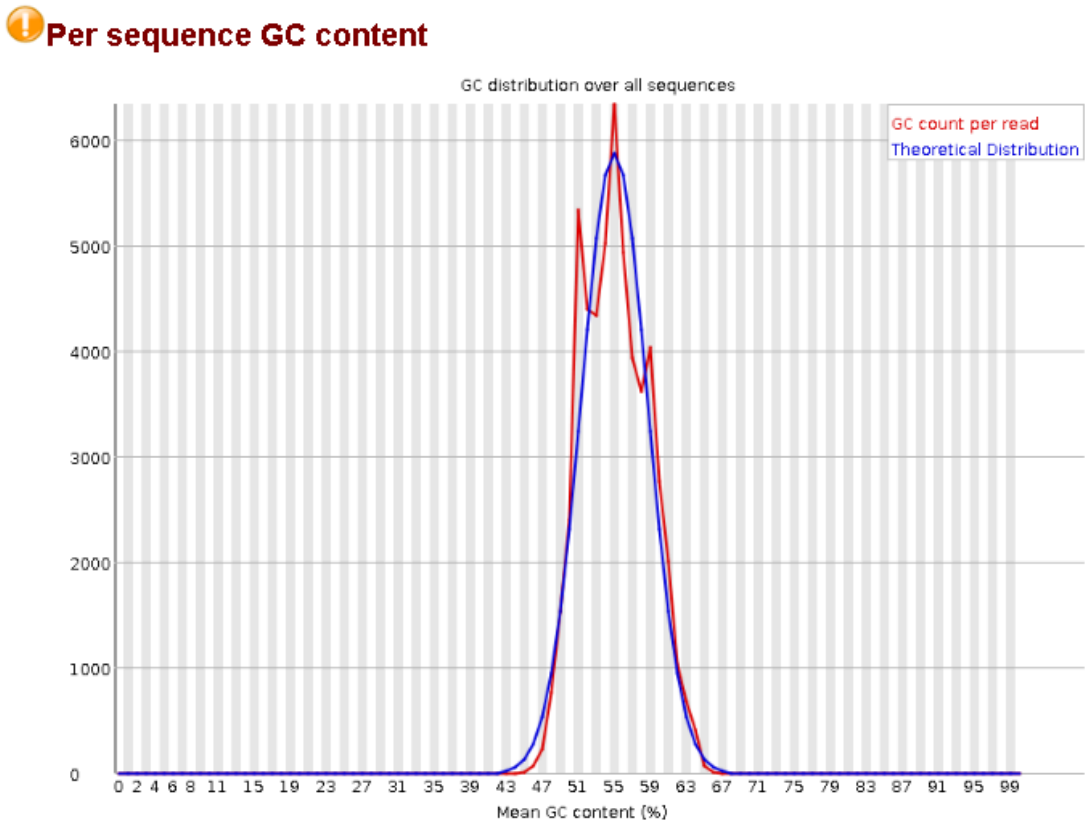


Рис. 17. **Вміст та розподіл GC для кожної послідовності**

Sequence Length Distribution. На цьому графіку показано розподіл розмірів фрагментів у проаналізованому файлі. У багатьох випадках це буде простий графік, що показуватиме пік лише одного розміру, але для файлів FASTQ змінної довжини це покаже відносні кількості кожного фрагмента послідовності різного розміру.

Цей модуль видасть попередження, якщо всі послідовності мають різну довжину.

Цей модуль викликає помилку, якщо будь-яка з послідовностей матиме нульову довжину.

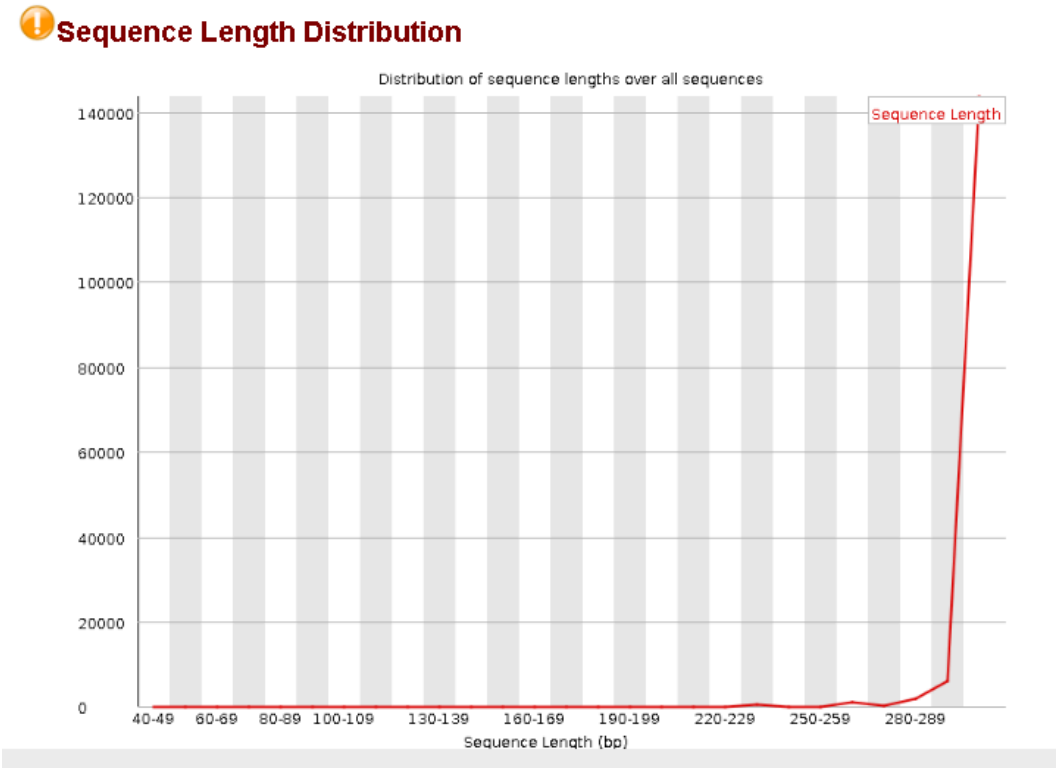


Рис. 18. **Розподіл розмірів фрагментів у проаналізованому читанні**

Sequence Duplication Levels. У різноманітній бібліотеці більшість послідовностей зустрічається лише один раз у фінальному наборі. Низький рівень дуплікації може вказувати на дуже високий рівень охоплення цільової послідовності, але високий рівень дуплікації, швидше за все, вказує на деяке зміщення збагачення (наприклад, ПЛР-ампліфікація).

Цей модуль підраховує ступінь дублювання для кожної послідовності в бібліотеці та створює графік, що показує відносну кількість послідовностей з різним ступенем дублювання.

Оскільки виявлення дублювання вимагає точного збігу послідовності по всій довжині послідовності, будь-які читання довжиною понад 75 п.н. усікаються до 50 п.н. для цілей цього аналізу. Тим не менш, більш довгі читання з більшою ймовірністю будуть містити помилки секвенування, які штучно збільшать розмаїтість, що спостерігається, і будуть мати тенденцію до недопредставлення сильно дубльованих послідовностей.

Цей модуль видасть попередження, якщо неунікальні послідовності становитимуть понад 20% від загальної кількості.

Цей модуль видасть помилку, якщо неунікальні послідовності становитимуть понад 50% від загальної кількості.

Як ми бачимо (рис. 19) наша послідовність містить послідовності що дублюються розміром більше 100 п.н. та розміром від 1к до 5к п.н.

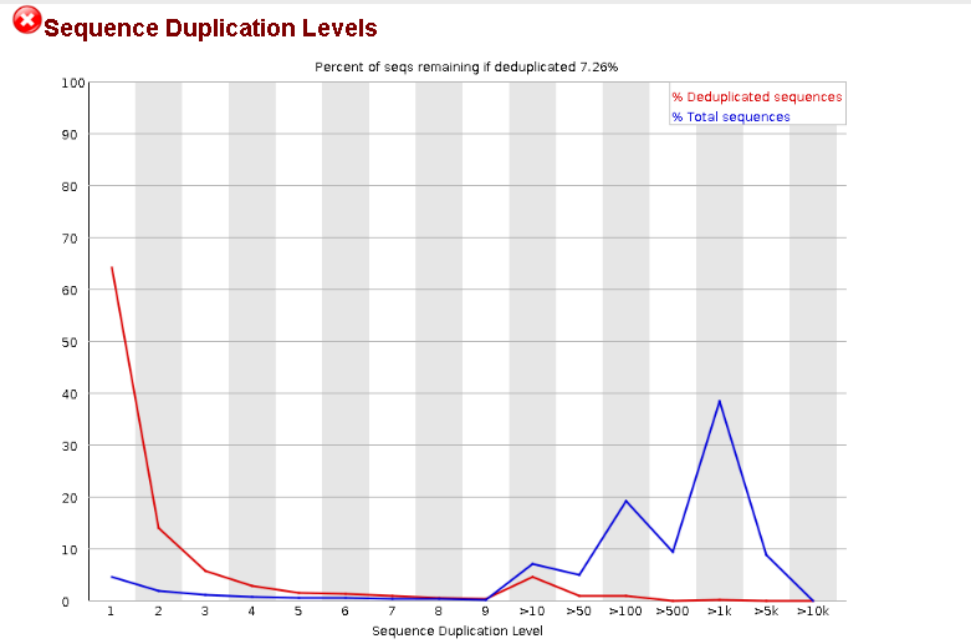


Рис. 19. **Графік наявності послідовностей, що повторюються**

Overrepresented sequences. Виявлення того, що одна послідовність надто широко представлена в наборі, означає або її високу біологічну значимість, або вказує на те, що бібліотека забруднена, або не така різноманітна, як ви очікували. У цьому модулі перераховані всі послідовності, які становлять понад 0,1% від загальної кількості. Для економії пам'яті до кінця файлу відстежуються ті послідовності, які входять у перші 100 000 послідовностей. Для кожної перепредставленої послідовності програма шукатиме збіги в базі даних загальних забруднювачів та повідомить про найкращий збіг, який вона знайде. Звернення повинні мати довжину не менше ніж 20 п.н. та мати не більше 1 невідповідності. Виявлення влучення не обов'язково означає, що це джерело зараження, але може вказати

правильний напрямок. Також варто відзначити, що багато послідовностей адаптерів дуже схожі один на одного, тому ви можете отримати повідомлення про влучення, яке технічно невірне, але має дуже схожу послідовність на фактичний збіг.

Оскільки виявлення дублювання вимагає точного збігу послідовності по всій довжині послідовності, будь-які читання довжиною понад 75 п.н. усікаються до 50 п.н. для цілей цього аналізу. Тим не менш, більш довгі читання з більшою ймовірністю будуть містити помилки секвенування, які штучно збільшать розмаїтість, що спостерігається, і будуть мати тенденцію до недопредставлення сильно дубльованих послідовностей.

Цей модуль видасть попередження, якщо буде виявлено, що будь-яка послідовність становить понад 0,1% від загального числа.

Цей модуль видасть помилку, якщо буде виявлено, що будь-яка послідовність становить понад 1% від загального числа.

Як ми бачимо, наша послідовність містить велику кількість перепредставлених послідовностей (рис. 20).

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CCTACGGGAGGCAGCAGTGAGGGATATTGGACAATGGATGGAAATCTGAT	7259	4.665167895115681	No Hit
CCTACGGGGGGCAGCAGTGAGGGATATTGGACAATGGATGGAAATCTGAT	6733	4.327120822622108	No Hit
CCTACGGGTGGCAGCAGTGAGGGATATTGGACAATGGATGGAAATCTGAT	4330	2.782776349614396	No Hit
CCTACGGGGGGCAGCAGTGAGGAATATTGCGCAATGGGCGAAAGCCTGAC	3078	1.9781491002570695	No Hit
CCTACGGGAGGCAGCAGTGAGGAATATTGCGCAATGGGCGAAAGCCTGAC	3049	1.9595115681233934	No Hit
CCTACGGGAGGCAGCAGTGGGGAATATTGGGCAATGGGCGCAAGCCTGAC	2655	1.7062982005141387	No Hit
CCTACGGGGGGCAGCAGTGGGGAATATTGGGCAATGGGCGCAAGCCTGAC	2627	1.688303419023137	No Hit
CCTACGGGAGGCAGCAGTGGGGAATATTGGGCAATGGGGGAAACCCTGAC	2445	1.57133676092545	No Hit
CCTACGGGGGGCAGCAGTGGGGAATATTGGGCAATGGGGGAAACCCTGAC	2382	1.530848329048843	No Hit
CCTACGGGAGGCAGCAGTGGGGAATATTGGGCAATGGGGGCAACCCTGAC	2182	1.4023136246786634	No Hit
CCTACGGGGGGCAGCAGTGGGGAATATTGGACAATGGGCGCAAGCCTGAT	2045	1.31426735218509	No Hit
CCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGCAAGCCTGAT	2023	1.3001285347043703	No Hit
CCTACGGGGGGCAGCAGTGGGGAATATTGGGCAATGGGGGCAACCCTGAC	1990	1.2789293084832905	No Hit
CCTACGGGAGGCAGCAGTGAGGAATATTGGACAATGGATGGAAATCTGAT	1988	1.2776349614395888	No Hit
CCTACGGGAGGCAGCAGTGAGGAATATTGGTCAATGGACGCAAGTCTGAA	1938	1.2455012853470437	No Hit
CCTACGGGTGGCAGCAGTGAGGAATATTGCGCAATGGGCGAAAGCCTGAC	1930	1.2403598971722365	No Hit
CCTACGGGGGGCAGCAGTGAGGAATATTGGTCAATGGACGCAAGTCTGAA	1914	1.2300771208226222	No Hit
CCTACGGGTGGCAGCAGTGGGGAATATTGGGCAATGGGCGCAAGCCTGAC	1759	1.1304627249357326	No Hit
CCTACGGGGGGCAGCAGTGAGGAATATTGGACAATGGATGGAAATCTGAT	1729	1.1111825192802056	No Hit
CCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGGGAAACCCTGAT	1709	1.0983298488431876	No Hit
CCTACGGGGGGCAGCAGTGGGGAATATTGGACAATGGGGGAAACCCTGAT	1635	1.050771208226221	No Hit

Рис. 20. **Графік виявлених перепредставлених послідовностей**

Adapter Content. Графік показує сукупний відсоток прочитань послідовностей адаптера кожної позиції. Як тільки послідовність адаптера виявляється в читанні, вона вважається присутньою до кінця читання, тому відсоток збільшується з довжиною читання. FastQC може визначати деякі стандартні адаптери (наприклад, Illumina, Nextera), для інших ми можемо надати файл забруднювачів як вхідні дані для інструмента FastQC.

В ідеалі, дані послідовності Illumina не повинні містити послідовності адаптера.

Цей модуль видасть попередження, якщо будь-яка послідовність є більш ніж у 5% всіх операцій читання.

Цей модуль видасть помилку, якщо будь-яка послідовність є більш ніж у 10% всіх операцій читання.

Як ми бачимо, наша послідовність не містить знайдених послідовностей адаптерів (рис. 21).

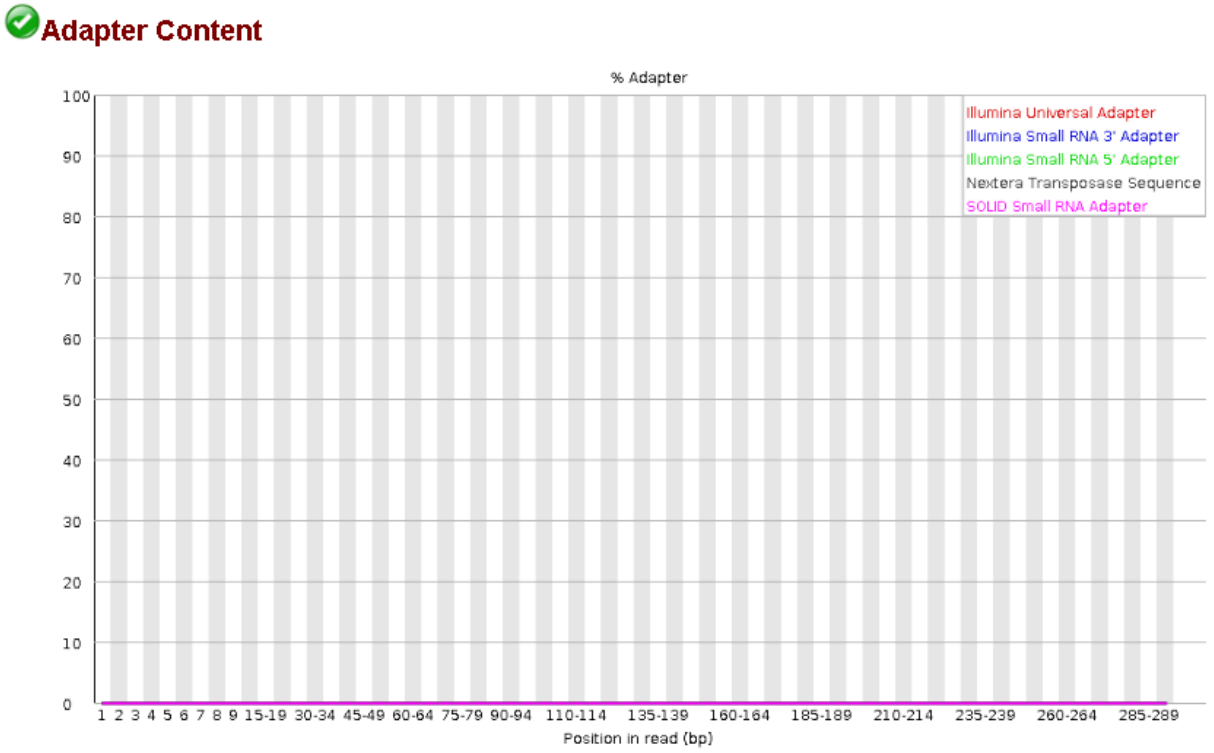


Рис. 21 Графік наявності адаптерів

Контрольні питання

1. Як завантажити послідовність в Galaxy?
2. Як контролювати якість даних NGS?
3. В якому форматі повинні бути завантажені послідовності для роботи в FastQC?
4. Які параметри якості слід перевіряти для кожного набору даних?
5. Як покращити якість набору даних послідовності?
6. Що може бути причиною збою у графіку вмісту базової послідовності?
7. Який модуль показує, що низька якість послідовності?
8. Який модуль каже вам, що існують надмірно подані (перепредставлені) послідовності? Про що нам це говорить?
9. У якому модулі можна побачити довжину читання набору контрольних даних?

СПИСОК ЛІТЕРАТУРИ

1. Мокроусов И. В. Методологические подходы к генотипированию *Mycobacterium tuberculosis* для эволюционных и эпидемиологических исследований. / И. В. Мокроусов //Инфекция и иммунитет. – 2012. – Т. 2, № 3. – С. 603–614.

2. Соломатина Е. В., Рахманов Р.С., Потехина Н.Н. Молекулярно-генетические методы в организации федерального государственного санитарно-эпидемиологического надзора за парентеральными гепатитами среди лиц опасных профессий. /Е. В. Соломатина, Р.С. Рахманов, Н.Н. Потехина //Медицинский альманах. – 2012. – Т. 22, № 3. – С. 90–93.

3. Andrews S. Fastqc. a quality control tool for high throughput sequence data / S. Andrews /citeulike– Електр. дан. – Режим доступу: http://www.bioinformatics.babraham.ac.uk ? /projects/fastqc/ (дата звернення: 23.03.2019). – Загол. з екрана.

4. Behjati S. What is next generation sequencing?/S. Behjati, P. S. Tarpey// Arch Dis Child Educ Pract Ed. – 2013. – Vol. 98(6). – P. 236–238.

5. Brown J. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool/ J. Brown, M. Pirrung, L. A. McCue// Bioinformatics. – 2017. – Vol.33, No. 19. – P. 3137–3139.

6. Campbell M. A Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis/ M. A Campbell, B. J. Haas, J. P. Hamilton, S. M. Mount, C. R. Buell// BMC Genomics. – 2008. – Vol. 7.–Articak: 327.

7. Cronin M. Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology/ M. Cronin, J. S. Ross // Biomark Med. – 2011. – Vol. 5, No. 3.– P. 293–305.

8. Desai A N. Next-generation sequencing: ready for the clinics? / A. N. Desai, A. Jere / /Clin Genet. – 2012. – Vol. 81. No. 6. – P. 503–510.

9. Gardy J.L. Whole-genome sequencing and social-network analysis of a tuberculosis outbrea. / J.L. Gardy, J.C. Johnston, S.J.H. Sui //N. Engl. J. Med. – 2011. – Vol. 364.– P. 730-739.

10. Grad Y.H. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. / Y.H. Grad, M. Lipsitch, M. Feldgarden //PNAS. – 2012. – Vol. 109, No. 8. – P. 3065–3070.

11. Hasan N.A.. Genomic diversity of 2010 Haitian cholera outbreak strains./ N.A. Hasan, S.Y. Choi, M. Eppinger // PNAS. – 2012. – Vol. 109, No. 29. – P. 210–217.

12. Hoff K. The effect of sequencing errors on metagenomic gene prediction / K. Hoff // BMC Genomics. – 2009. – Vol. 10. – P. 520–561.

13. Ku C S. Technological advances in DNA sequence enrichment and sequencing for germline genetic diagnosis. / C. S. Ku, M. Wu, D. N. Cooper // Expert Rev Mol Diagn. – 2012. – Vol. 12, No. 2. – P. 159–173.

14. Koser C.U. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology /C.U. Koser, M.J. Ellington, E.J.P. Cartwright //PLoS Pathogen. – 2012. – Vol. 8, No. 8. – P. 1–9.

15. Maxam A.M. A new method of sequencing DNA. / A.M. Maxam, W. Gilbert // PNAS. – 1977. – Vol. 74.– P. 560-564.

16. Meldrum C. Next-generation sequencing for cancer diagnostics: a practical perspective/ C. Meldrum, M. A. Doyle, R.W. Tothill // Clin Biochem Rev. – 2011. – Vol. 32, No. 4. – P. 177–195.

17. Mitra R.D. *In situ* localized amplification and contact replication of many individual DNA molecules . / R.D. Mitra, G.M. Church //Nucleic Acids Res. – 1999. – Vol. 27, No. 24. – e34.

18. Moorthie S. Review of massively parallel DNA sequencing technologies. /S. Moorthie, C.J. Mattocks, C.F. Wright //Hugo J. – 2011. – No. 5. – P. 1–12.

19. Nielsen R. Genomics: In search of rare human variants/ R. Nielsen / // Nature. – 2010. – Vol.467, No. 7319. – P. 1050–1051.

20. Nilesh T. H. Next-generation sequencing technologies: An overview/ T. H. Nilesh, C. D. Monosa, A. Dinh// Human Immunology. – 2021. – Vol. 82, Issue 11. – P. 801-811

21. Pareek C.S. Sequencing technologies and genome sequencing. / C.S. Pareek , R. Smoczynski, A. Tretyn. //J. Appl. Genetics. 2011. Vol. 52. № 4. P. 413–435.

22. Pettersson E. Generations of sequencing technologies./ E. Pettersson, J. Lundeberg, A. Ahmadian. // Genomics. – 2009. – Vol. 93, No. 2. – P. 105–111.

23. Quail M. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers/ M. Quail, M. Smith, P Coupland // BMC Genomics. – 2012. – Vol. 13. – P. 341–351.

24. Sanger F. Coulson A.R. DNA sequencing with chain terminating inhibitors. / F. Sanger , S. Nicklen //PNAS. – 1977. – Vol. 74, No. 12. – P. 5463–5467.

25. Slatko B. E. Overview of Next Generation Sequencing Technologies/ B. E. Slatko, A. F. Gardner, F. M. Ausubel// Curr Protoc Mol Biol. – 2018. –Vol. 122(1). – Aryical: e59.

26. Taub M. Overcoming bias and systematic errors in next generation sequencing data / M. Taub, H. Bravo, R. Irizarry // Genome Medicine. – 2010. – Vol.2. – P. 192–198.

27. Voelkerding K.V. Next-generation sequencing: from basic research to diagnostics. / K.V. Voelkerding, S.A. Dames, J.D. Durtschi //Clin. Chem. – 2009. – Vol. 55, No 4. – P. 641–658.

28. Zhang J., The impact of next-generation sequencing on genomics./ J. Zhang, R. Chiodini, A. Badr// J. Genet. Genomics. – 2011. – Vol. 38, No. 3. – P. 95–109.