

UNIWERSYTET WROCŁAWSKI

WYDZIAŁ BIOTECHNOLOGII

ZAKŁAD MIKROBIOLOGII MOLEKULARNEJ

Przygotowanie bioinformatycznej bazy danych NGS
dotyczących bakterii z rodzaju *Streptomyces*

Marek Skrzyński

Praca magisterska

Opiekun pracy: dr Agnieszka Strzałka

Wrocław, 2025 r.

Podziękowania

Najserdeczniejsze podziękowania składam mojej Pani Promotor, dr Agnieszce Strzałce, za poszerzenie mojej wiedzy bioinformatycznej, a przede wszystkim za okazaną wyrozumiałość i cierpliwość na każdym etapie naszej współpracy.

Z głębi serca dziękuję moim Rodzicom za wsparcie, które okazują mi od kiedy tylko pamiętam, za wiarę we mnie oraz miłość, którą nieustannie od nich otrzymuję.

Dziękuję Magdzie za to, że zawsze jest przy mnie i za motywację do działania w trudniejszych chwilach.

Streszczenie

Celem tej pracy magisterskiej było stworzenie narzędzia bioinformatycznego umożliwiającego analizę danych transkryptomicznych bakterii z rodzaju *Streptomyces*, dokładnie *Streptomyces coelicolor* i *Streptomyces venezuelae*. Bakterie te są organizmami modelowymi, wykazują złożony cykl życiowy i są źródłem wielu metabolitów drugorzędowych, w tym wielu antybiotyków. Wiele badań skupia się właśnie na aspektach powstawania metabolitów drugorzędowych, a dokładniej na analizie i optymalizacji procesów zaangażowanych w ich powstawanie. Wynika to z faktu, że kontrola ekspresji genów tych gatunków jest bardzo rozbudowana ze względu na rozwinięty cykl życiowy i zmienne środowisko. Stąd zapotrzebowanie na zgromadzenie w jednym miejscu danych dotyczących tych eksperymentów, aby badacze mogli mieć do nich szybki wgląd i w łatwy sposób porównać z nimi swoje wyniki. Dlatego na podstawie zebranych danych eksperymentalnych z różnych publikacji naukowych, zaprojektowana i stworzona została aplikacja „Strep.N.A”. Jest to interaktywna baza danych pozwalająca na analizę i wizualizację wyników. Program powstał w języku R z wykorzystaniem pakietu Shiny. Dostępne dla użytkowników jest porównywanie obecnych w bazie danych, załadowanie własnych danych, filtrowanie wyników na podstawie niektórych parametrów, jak i pobranie powstałych wizualizacji. Narzędzie to ma na celu wspierać badania nad regulacją ekspresji genów i produkcją metabolitów drugorzędowych bakterii *Streptomyces*.

Abstract

The aim of this master's project was to create a bioinformatics tool for the analysis of transcriptomic data from bacteria of *Streptomyces* genus, specifically *Streptomyces coelicolor* and *Streptomyces venezuelae*. These model representatives of this genus exhibit a complex life cycle and thanks to this, they produce a lot of secondary metabolites, including many antibiotics. Numerous studies focus on the processes involved in biosynthesis of these compounds, particularly, in their analysis and optimization. The gene expression regulation is very complex in this *Streptomyces* species, mostly because of their developed life cycle, and changes in the environment. That's why there is a need to collect experimental data from such studies in one data base, providing researchers a quick access and ability to easily compare their own results. Based mainly on gathered experimental data, and what we wanted to achieve with this program, the interactive application „Strep.N.A” was designed and developed. It functions as a database that enables the analysis and visualization of transcriptomic results. The app was created with R programming language, and the Shiny package. It allows users to compare data available in the database, upload their own datasets, filter results based on selected parameters and download created plots in image format. Our app is intended to support research on gene expression regulation and secondary metabolite production in *Streptomyces* bacteria.

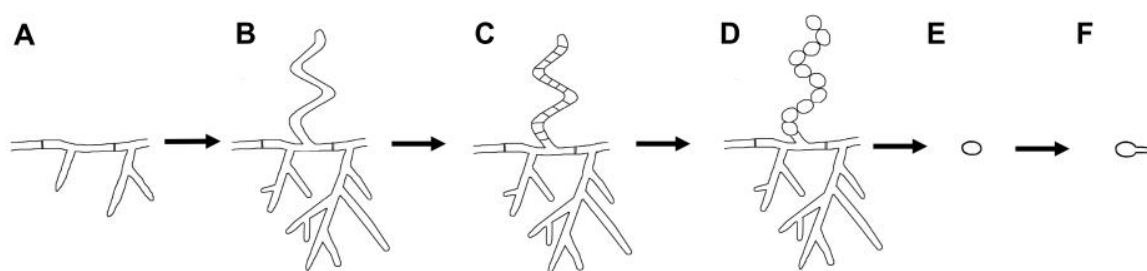
Spis treści

Streszczenie	3
Abstract	4
1. Wstęp.....	6
1.1 <i>Streptomyces</i>	6
1.2 Gatunki modelowe <i>Streptomyces</i>	7
1.3 Sekwencjonowanie nowej generacji.....	12
1.4 Bazy danych i narzędzia do analizy danych sekwencjonowania	18
2. Cel pracy	24
3. Materiały i metody	25
4. Wyniki	37
4.1 Program Strep.N.A	37
4.1.1 Podstawowe informacje po włączeniu	38
4.1.2 Panel boczny i funkcje „RNA-seq & ChIP-seq visualization”	40
4.1.3 Wyświetlanie danych „RNA-seq & ChIP-seq visualization”	43
4.1.4 Panel boczny i funkcje „Comparison of data”	46
4.1.5 Wyświetlanie danych „Comparison of data”	49
4.1.6 Strona „Help&Info”	52
4.2 Zastosowanie aplikacji	53
4.2.1 RNA-seq i ChIP-seq	53
4.2.2 RNA-seq i RNA-seq	56
4.2.3 Poznanie funkcji nieznanego białka na podstawie zmian ekspresji.....	60
5. Dyskusja	62
6. Bibliografia.....	66
7. Skróty i sformułowania anglojęzyczne	75

1. Wstęp

1.1 *Streptomyces*

Promieniowce, obszerny rząd organizmów prokariotycznych, należą do bakterii Gram-dodatnich. Grupę tą cechuje wysoka zawartość par GC w materiale genetycznym, rozbudowany metabolizm drugorzędowy i wysoki poziom zróżnicowania, często w obrębie jednego gatunku pod kątem morfologicznym [1]. Do Promieniowców należą, między innymi, bakterie rodzaju *Streptomyces*. *Streptomyces* to bardzo obszerny rodzaj bakterii liczący ponad 1000 gatunków. Posiadają większy genom niż przeciętni reprezentanci prokariota, gdyż u *Streptomyces coelicolor* to aż 8,7 miliona par zasad, i tak jak u innych reprezentantów Promieniowców, cechuje je wysoka zawartość par GC [2], [3]. Są to bakterie znajdowane w glebie, jak i w morzach czy oceanach, szczególnie na dnie morskim i wodorostach. Większość z przedstawicieli tego rodzaju charakteryzuje złożony cykl życiowy składający się z fazy wzrostu wegetatywnego i sporulacji. Sposób rozmnażania się *Streptomyces* przyrównywany jest do tego u grzybów strzępkowych, lecz powstające u nich strzępko-podobne twory są mniej złożone morfologicznie i o mniejszym rozmiarze [4]. Umiejętność tworzenia strzępek wegetatywnych pozwala im na docieranie do składników odżywczych potrzebnych do ich przetrwania [5] [Rys. 1.1]. Jednakże, gdy składników w okolicy bakterii zabraknie, wytwarzają one grzybnie powietrzne. W wyniku podziału strzępek powietrznych powstają zarodniki, dzięki czemu bakterie mogą szybko rozprzestrzenić się na nowym terenie bogatym w składniki odżywcze i uniknąć niekorzystnych dla siebie warunków [6].



Rysunek 1.1 Cykl życiowy *Streptomyces*. Schemat głównych etapów wzrostu: (A) Strzępki wegetatywne, (B) Wczesna sporulacja rozpoczynająca się od formowania grzybni powietrznej, (C) Powstanie niedojrzałych zarodników, (D) Dojrzewanie zarodników, (E) Uwalnianie dojrzałych zarodników, (F) Kielkowanie strzępki wegetatywnej. Zaadaptowano z [7].

W wyniku tak rozbudowanych faz wzrostu powstaje wiele metabolitów drugorzędowych, które nie są niezbędne dla *Streptomyces* do wzrostu i przetrwania. Dzięki tej właściwości, ów rodzaj bakterii postrzegany jest jako źródło gamy różnorodnych związków naturalnego pochodzenia. Produkują one związki antybiotyczne takie jak streptomycyna, daptomycyna czy awermektyna [8]. Przy użyciu *Streptomyces* wyprodukowano również związki działające przeciw grzybom, stosowane nawet przy infekcjach grzybiczych, takie jak amfoterycyna [8]. Przeprowadza się też badania nad klastrami genów, które potencjalnie mogłyby przyczyniać się do produkcji metabolitów drugorzędowych, lecz w warunkach laboratoryjnych są nieaktywne [9]. Również tematem ostatnio rozpatrywanym jest zastosowanie *Streptomyces* w heterologicznej produkcji białek, z nadzieją na lepszą wydajność niż u *Escherichia coli* [10].

1.2 Gatunki modelowe *Streptomyces*

Główne dwa gatunki *Streptomyces*, szeroko stosowane w badaniach biotechnologicznych, to *Streptomyces coelicolor* i *Streptomyces venezuelae*. *S. coelicolor* możemy nazwać klasycznym organizmem modelowym, ponieważ był to pierwszy gatunek *Streptomyces*, którego genom został w całości zsekwencjonowany, natomiast *S. venezuelae* jest nowym modelem ze względu na wyróżniającą się szybkość wzrostu i charakterystykę wzrostu w podłożu płynnym [2], [11]. *S. coelicolor* posiada około 7,8 tysiąca genów, a w całym genomie około 27 klastrów genów biosyntezy (*biosynthetic gene clusters*, BGC) kodujących białka, umożliwiające produkcję wyspecjalizowanych metabolitów. Bakteria ta szybko zyskała zainteresowanie w oczach badaczy ze względu na kolor kolonii, które tworzy ten gatunek. *S. coelicolor* wytwarza związek zwany aktynorodyną, o kolorze niebieskim [Rys. 1.2], który jest antybiotykiem kodowanym przez klaster genów *act*, a do tego jego barwa jest zależna od pH hodowli [11], [12]. *S. coelicolor* wytwarza również inne pigmenty w kolorach czerwonym, żółtym i szarym. Dzięki tej naturalnej właściwości, *S. coelicolor*, łatwo można odróżnić mutanty zablokowane na różnych etapach szlaku biosyntezy [11].



Rysunek 1.2 Zdjęcie kolonii *S. coelicolor* z grzybnią powietrzną i zabarwionymi na szaro zarodnikami. Ciemne zabarwienie wokół kolonii i krople na kolonii są wydzieloną aktynorodyną. Zaadaptowano z [11].

S. coelicolor jako reprezentant *Streptomyces*, posiada rozbudowany cykl życiowy, dlatego bada się regulację ekspresji genów i mechanizmów kontrolujących transkrypcję, w tym na przykład mutacje delecji genów związanych z rozwojem strzępek i fazy sporulacji. Jednym z najważniejszych mechanizmów regulacji transkrypcji u *Streptomyces* jest kontrola przez czynniki sigma. Czynniki sigma wchodzi w skład bakteryjnych polimeraz RNA i decydują one o specyficzności promotora, a w przypadku *Streptomyces* odgrywają również rolę w przekazywaniu sygnałów co ułatwia im przetrwanie w trudnych warunkach [13]. U *S. coelicolor* znajduje się około 64 czynników sigma, odgrywających różne role w cyklu życiowym tego organizmu. Można je podzielić na 4 grupy, różniące się ze względu na swoje funkcjonowanie i budowę [13]. Grupę pierwszą cechuje zdolność regulacji transkrypcji genów niezbędnych do funkcjonowania komórki, stąd tą grupę nazywa się konstytutywnymi czynnikami sigma. Grupa druga choć bardzo podobna w budowie do grupy pierwszej, odpowiada za regulację genów nieistotnych do wzrostu bakterii. Grupa trzecia, określana również jako alternatywne czynniki sigma, reguluje wiele szlaków metabolicznych w odpowiedzi na sygnały wewnątrz- jak i zewnątrzkomórkowe. Natomiast grupa czwarta odgrywa rolę w regulacji odpowiedzi na stres [13]. Czynniki sigma są niezwykle istotne w regulacji transkrypcji i funkcjonowania całej komórki, stąd przeprowadza się

doświadczenia polegające między innymi na usuwaniu genów kodujących czynniki sigma. Usunięcie czynnika sigma B, należącego do grupy trzeciej wykazało, że jest on głównym regulatorem, który aktywuje inne czynniki sigma w kaskadzie, a tym samym kontroluje proces różnicowania się komórek oraz odpowiedź oksydacyjną *S. coelicolor* [14].

Streptomyces posiadają również setki czynników transkrypcyjnych, innych niż czynniki sigma, które modułują ekspresję genów na poziomie globalnym komórki, jak i takich, które odpowiadają za specyficzne szlaki biosyntezy. Globalne czynniki transkrypcyjne zwykle regulują ekspresję nawet całego klastra genów, w którym mogą się znajdować [15]. Do takich należą między innymi geny *bld*. Mutanty *bld*, bez genów regulatorowych *bld*, które odpowiedzialne są za powstawanie strzępek powietrznych, nie posiadają na koloniach „włosków” typowych dla kolonii *S. coelicolor*, stąd z ang. „bald” otrzymuje się fenotyp „łysy”, łatwy do rozpoznania. Innym czynnikiem działającym szeroko na regulację procesów komórkowych jest grupa genów *whi*, odpowiadających za podział strzępek powietrznych w spory. W mutantach *whi S. coelicolor*, nie zauważymy szarego pigmentu charakterystycznego dla dojrzałych spor. Z kolei do genów o małym zasięgu regulacji należą, między innymi, ActIIORF4, czy RedD [15], które odpowiadają jedynie za regulację biosyntezy konkretnych metabolitów wtórnych występujących w *S. coelicolor*. Jednak poszerzanie wiedzy powoduje, że tak proste zależności są obalane. Czasem dochodzi do regulacji krzyżującej się między różnymi genami i spotkać się możemy z sytuacją, w której gen regulujący specyficzny szlak, reguluje również gen o szerszym zakresie działania [15]. Podobnym czynnikiem regulującym globalnie jest AdpA. Koordynuje on wiele procesów komórkowych, jak różnicowanie morfologiczne, replikacja chromosomów jak i syntezę metabolitów drugorzędowych, w tym syntezę chloramfenikolu u *S. venezuelae* [16].

Innym czynnikiem regulującym ekspresję genów są białka związane z organizacją nukleoidu (*Nucleoid-associated proteins*, NAP). Odgrywają one kluczową rolę w rearanżacji DNA, wiązaniu go i stabilizowaniu domen powstałych z materiału genetycznego [17]. Te funkcje NAP powodują, że wpływają one na dostępność promotorów dla czynników transkrypcyjnych i polimerazy RNA. Do NAP należą między innymi, białka z grupy homologów HU, HupA i HupS [17]. Są to białka, które wpływają na organizację DNA i wpływają na wzrost *S. coelicolor*. Mutanty z usuniętymi genami *hupA* i *hupS* wykazują zahamowanie wzrostu, defekty w sporulacji oraz większą podatność na czynniki stresowe [18]. Zaobserwowano również, że NAP mogą przysłużyć się do aktywacji wyciszonych klastrów genów biosyntetycznych, które domyślnie nie są aktywne [19].

Dzięki wszystkim powyższym właściwościom *S. coelicolor* były używane jako organizm modelowy dla *Streptomyces* przez lata. Niestety ma on swoje ograniczenia. Między innymi, jak wiele innych gatunków tego rodzaju, do jego sporulacji dochodzi tylko na podłożu stałym do tego w sposób asynchroniczny, a w mediach płynnych tworzy duże skupiska strzępek mycelialnych [11]. Wynikiem tego jest coraz większe zainteresowanie *S. venezuelae* jako reprezentantem *Streptomyces*. Jest to gatunek o trochę mniejszym genomie niż *S. coelicolor*, posiadającym 7,1 tysiąca genów kodujących białka, a w całym genomie około 34 BGC. Około 85% genów kodujących białka to ortologi genów *S. coelicolor*, ale w przeciwieństwie do *S. coelicolor*, *S. venezuelae* szybciej i łatwiej osiąga synchronizację rozwoju w mediach płynnych oraz przechodzi w nich cały cykl życiowy. Oprócz podstawowych faz wzrostu takich jak wegetatywny i sporulacyjny *S. venezuelae* wykazuje też tryb wzrostu eksploracyjnego, charakteryzującego się dynamicznym rozprzestrzenianiem się strzępek bakterii po podłożu, na przykład, w wyniku niedoboru składników odżywczych. W ten sposób bakteria może szybko rozprzestrzenić się w pobliskim środowisku bez konieczności sporulacji [20]. Dzięki tym cechom *S. venezuelae* jest przewodnim gatunkiem w badaniu cyklu życiowego bakterii *Streptomyces*, ponieważ z hodowli płynnych z łatwością można wykonać preparaty do badań mikroskopowych w tym do badań z użyciem mikroskopii fluorescencyjnej [21]. Charakterystycznym pigmentem dla *S. venezuelae* jest zielony barwnik powstający w dojrzałych zarodnikach [Rys. 1.3], [21]. Produkują one również inne związki, które są nieobecne u *S. coelicolor*, między innymi chloramfenikol, który jest antybiotykiem o bardzo szerokim działaniu, powodującym inhibicję syntezy białek poprzez wiązanie z podjednostką 50S rybosomu bakteryjnego [22]. Do tego większość metod stosowanych do modyfikacji genetycznych *S. coelicolor*, działa również na *S. venezuelae*

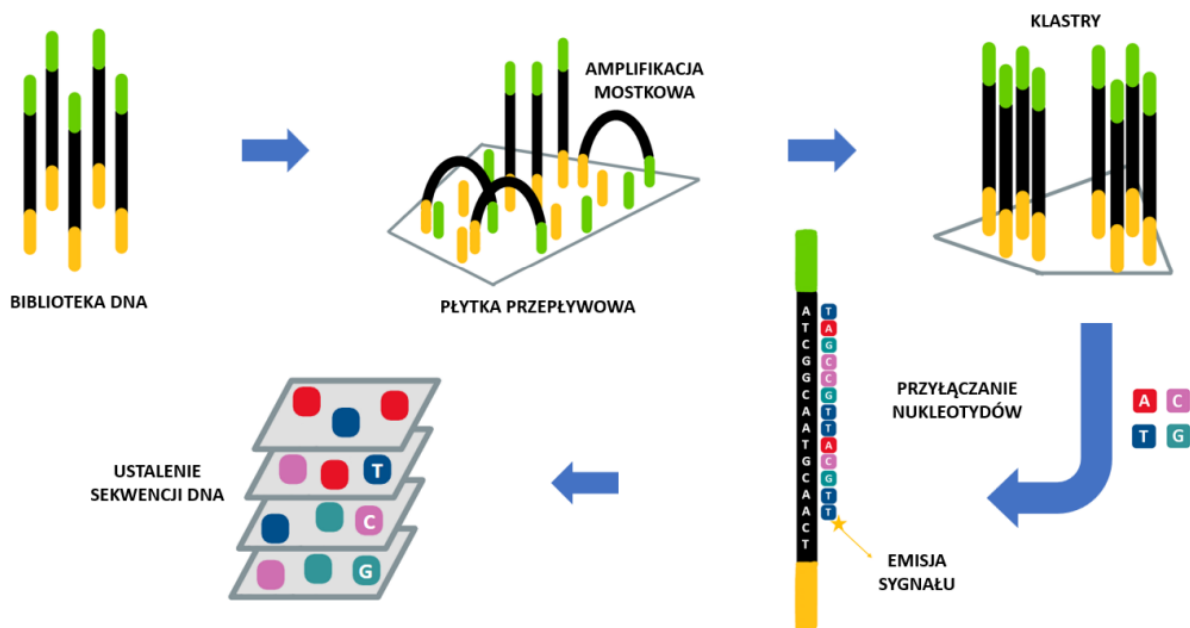


Rysunek 1.3 Kolonia *S. venezuelae* podczas sporulacji. Charakterystyczne zielone pokrycie spowodowane obecnością pigmentu o tym kolorze w dojrzałych zarodnikach [21].

Oba te gatunki cechuje rozbudowany cykl życiowy, wiele szlaków biosyntezy metabolitów drugorzędowych i klastrów biosyntezy genów. Zatem z tych cech wynika potencjał do usprawnienia wydajności produkcji antybiotyków, jak i ich samych, poprzez manipulacje BGC, które produkują naturalnie antybiotyki, oraz inżynierię szlaków metabolicznych i syntezę białek rekombinowanych pozwalających na nowe terapie. Od kilku lat kładzie się nacisk na zgłębianie działania szlaków biosyntezy w *Streptomyces*, a dokładniej, jak konkretnie wpływają one na siebie nawzajem i jakie modyfikacje możemy wprowadzać, aby usprawnić ich funkcjonowanie, bądź produkcję niektórych metabolitów. Oprócz tego bada się jak powiązane są te szlaki z procesami towarzyszącymi fazom wzrostu w warunkach hodowlanych, często pod wpływem różnego rodzaju czynników stresowych. Do takich rodzajów badań można stosować metody z rodzaju technik sekwencjonowania nowej generacji takie jak RNA-seq, ChIP-seq jak i techniki wykorzystującą hybrydyzację DNA zwaną mikromacierzami.

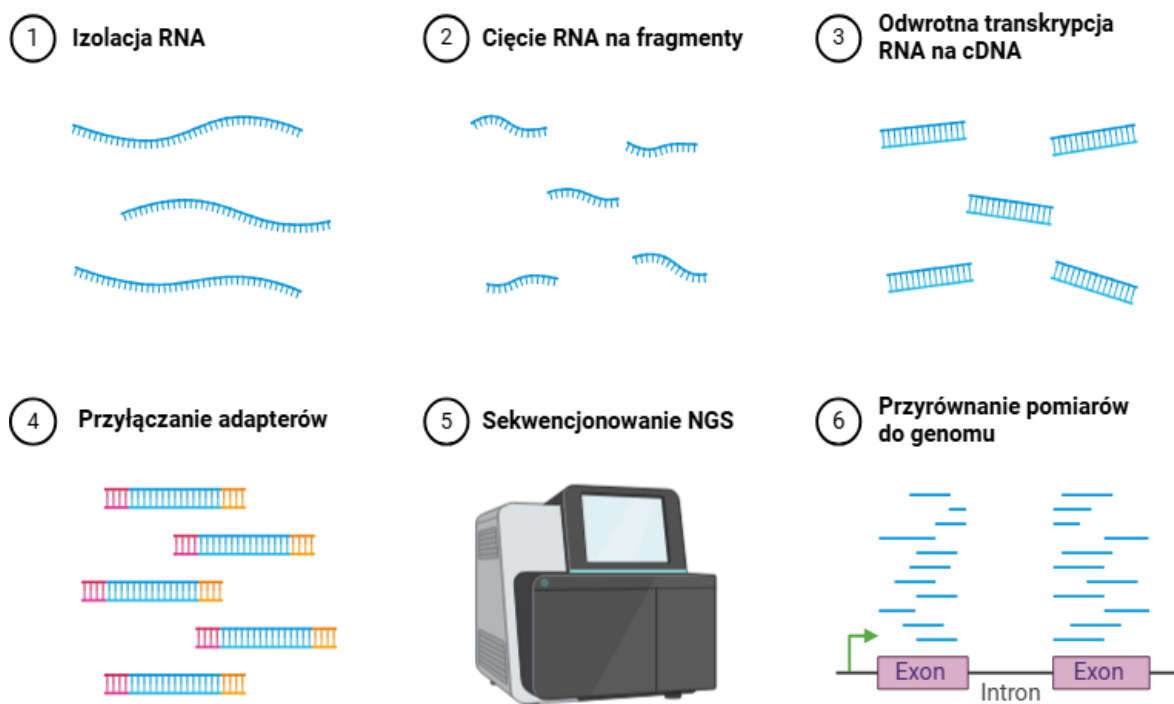
1.3 Sekwencjonowanie nowej generacji

Od zsekwencjonowania genomu ludzkiego techniki odczytywania sekwencji materiału genetycznego rozwijały się w gwałtownym tempie. Do zapisania jak wygląda genom ludzi zastosowane zostało sekwencjonowanie pierwszej generacji, a dokładnie sekwencjonowanie metodą Sangera. Największym problemem tej techniki na początku pod koniec XX wieku była jej niska przepustowość. Stąd pojawiła się potrzeba powstania nowych technik, które oprócz umożliwiania wyższej przepustowości badań, byłyby też bardziej opłacalne od strony finansowej. Dlatego już wtedy badacze zaczęli tworzyć techniki drugiej generacji sekwencjonowania, zwane teraz sekwencjonowaniem nowej generacji (*New generation sequencing*, NGS). Sekwencjonowanie pierwszej i drugiej generacji działają na podobnych zasadach, lecz NGS zdecydowanie kładzie nacisk na odczytanie dużo większej liczby odczytów w jednym momencie. Do tego metoda Sangera umożliwiała odczytanie sekwencji jednej nici, a metody NGS pozwalają na przeanalizowanie całego genomu na raz, w podobnej ilości czasu. Ponadto umożliwiają odczyt bardzo krótkich fragmentów, do 150 par zasad. Przede wszystkim w NGS używa się nukleotydów oznakowanych znacznikami fluorescencyjnymi, o 4 różnych kolorach odpowiadających zasadom azotowym A, C, T i G. Wszystkie nukleotydy używane w NGS są oznakowane w odróżnieniu od metody Sangera [23]. Do przeprowadzenia NGS przygotowuje się preparat DNA, oczyszczony z RNA i rRNA, a następnie DNA jest cięte na krótkie fragmenty. Następnie do fragmentów przyłączane są adaptory. Fragmenty DNA są rozdzielane na pojedyncze nici. Następnie próbkę przenosi się na kasetę przepływową, na której szklanej powierzchni znajdują się krótkie nukleotydy komplementarne do adapterów na fragmentach DNA. Pojedyncza nić DNA przyłącza się do oligonukleotydu na kasecie. Dochodzi wtedy do amplifikacji fragmentu na kasecie, powstają skupiska jednej matrycy, co zwiększa liczbę odczytów. Do fragmentów dołączane są nukleotydy fluorescencyjne [Rys. 1.4]. Między każdym dołączeniem nukleotydu dokonuje się pomiaru fluorescencji i zahamowania emisji przyłączonego nukleotydu. W ten sposób powstaje szereg kolejnych odczytów, z których jesteśmy w stanie odczytać jaka jest sekwencja danego fragmentu. Dzięki temu możliwa jest analiza milionów par zasad podczas jednego przebiegu NGS [24].



Rysunek 1.4 Schemat sekwencjonowania genetycznego nowej generacji [24].

Sekwencjonowanie RNA (RNA-seq), umożliwia analizę transkryptomu, czyli dokładniej poziomu ekspresji wszystkich genów w różnych warunkach, takich jak, różne fazy wzrostu czy obecność czynników stresowych. Istnieją różne systemy dostępne na rynku, takie jak FLX, Illumina czy AB SOLiD system, umożliwiające wykonywanie pomiarów RNA-seq, lecz operują one na podobnych założeniach [25]. Systemy te różnią się głównie tym jakiej długości fragmenty DNA są przyjmowane do sekwencjonowania. Głównym założeniem RNA-seq jest pomiar ekspresji, więc nie operuje się tu na próbkach DNA. Z kilku powtórzeń biologicznych pobierany jest materiał genetyczny RNA, który powstaje w wyniku ekspresji DNA obecnego w bakterii. RNA jest dzielone na fragmenty, które są transkrybowane na dwuniciowe cDNA i dopiero wtedy przeprowadza się ich sekwencjonowanie [Rys. 1.5]. Pozyskane krótkie fragmenty są sekwencjonowane równolegle, w efekcie czego otrzymujemy duże ilości krótkich odczytów [25]. Odczyty danych fragmentów są następnie zliczane i przyrównywane do genomu. W ten sposób można odczytać poziom ekspresji danego genu [Rys. 1.5]

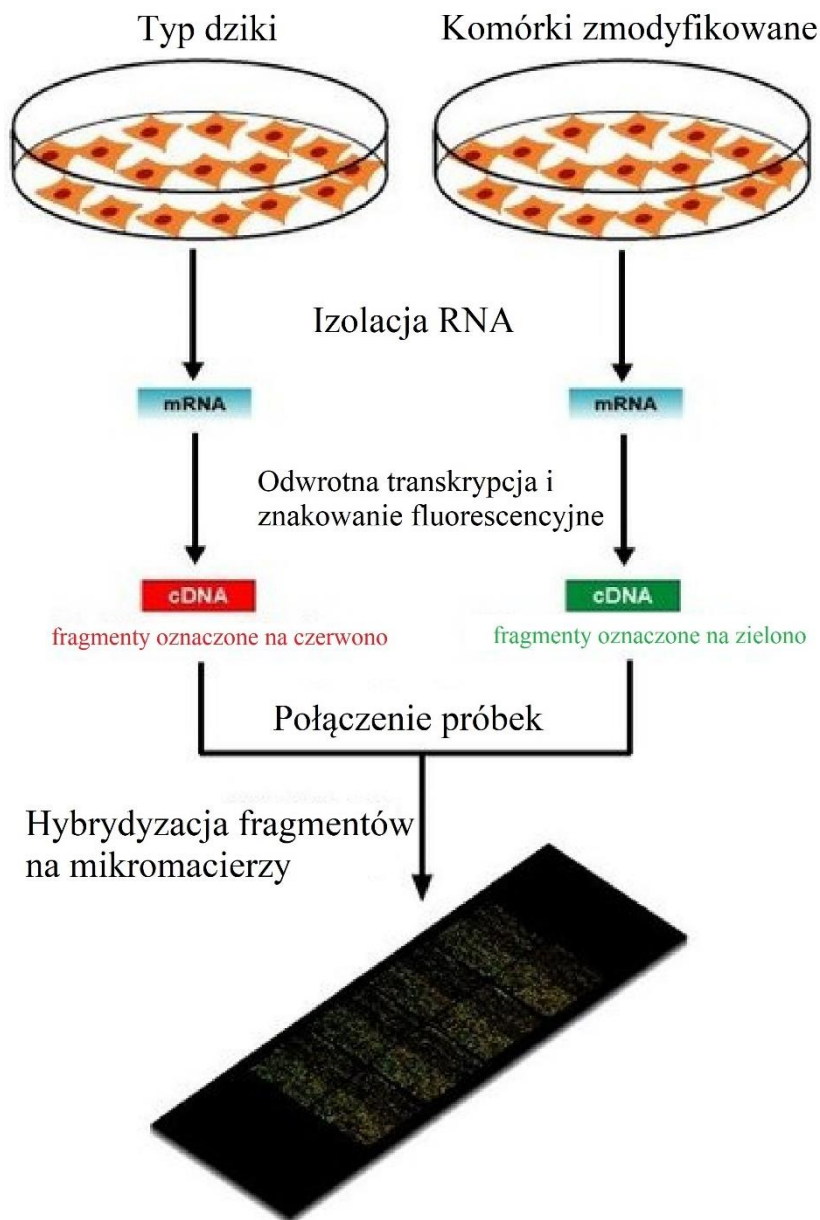


Rysunek 1.5 Schemat zasady działania RNA-seq. Zaadaptowano z [26]

Stosując metodę RNA-seq można dokonać pomiaru nie znając genomu wyjściowego. Jednak genom *S. coelicolor* i *S. venezuelae* jest znany. Przy sekwencjonowaniu genomu tych bakterii celem jest zaobserwowanie liczby odczytów danych genów, ponieważ dzięki takiemu podejściu można otrzymać różnice w ekspresji badanych przez nas genów. Zwykle ta różnica jest określana jako fold change (FC), czyli stosunek ekspresji genu w dwóch warunkach na przykład między typem dzikim bakterii i mutantem z delecją konkretnego genu. Parametr FC często podawany jest w postaci logarytmu o podstawie dwa i zapisywany jako „logFC”. Ważnym parametrem też jest wskaźnik fałszywych odczytów (*false discovery rate*, FDR). FDR liczony jest na podstawie wartości „p” dla pomiarów i dostarcza informacji o tym jaka część wyników istotnych statystycznie może być fałszywie pozytywna. Dlatego często analizę ogranicza się do wyników, dla których wartość FDR jest mniejsza od 0,05. RNA-seq można zastosować do porównania jak delecja danego genu wpłynie na geny z nim sąsiadujące, bądź na inne powiązane z nim geny w porównaniu do typu dzikiego tej samej bakterii [27]. Przykładem zastosowanie RNA-seq w badaniu *Streptomyces* może być zbadanie wpływu usunięcia genu *ohkA* odpowiedzialnego za kodowanie sierocej kinazy histydynowej, która pełni rolę w różnicowaniu morfologicznych. Brak *ohkA* spowodował zwiększenie ekspresji w klastrach genów odpowiedzialnych między innymi za karboksylazę acetylo-CoA, jak i biosyntezę aktynorodyny [27], [28]. Można też przeprowadzić pomiary w różnych fazach

wzrostu bakterii i porównać, jak wygląda ekspresja genów w różnych etapach cyklu życiowego [29].

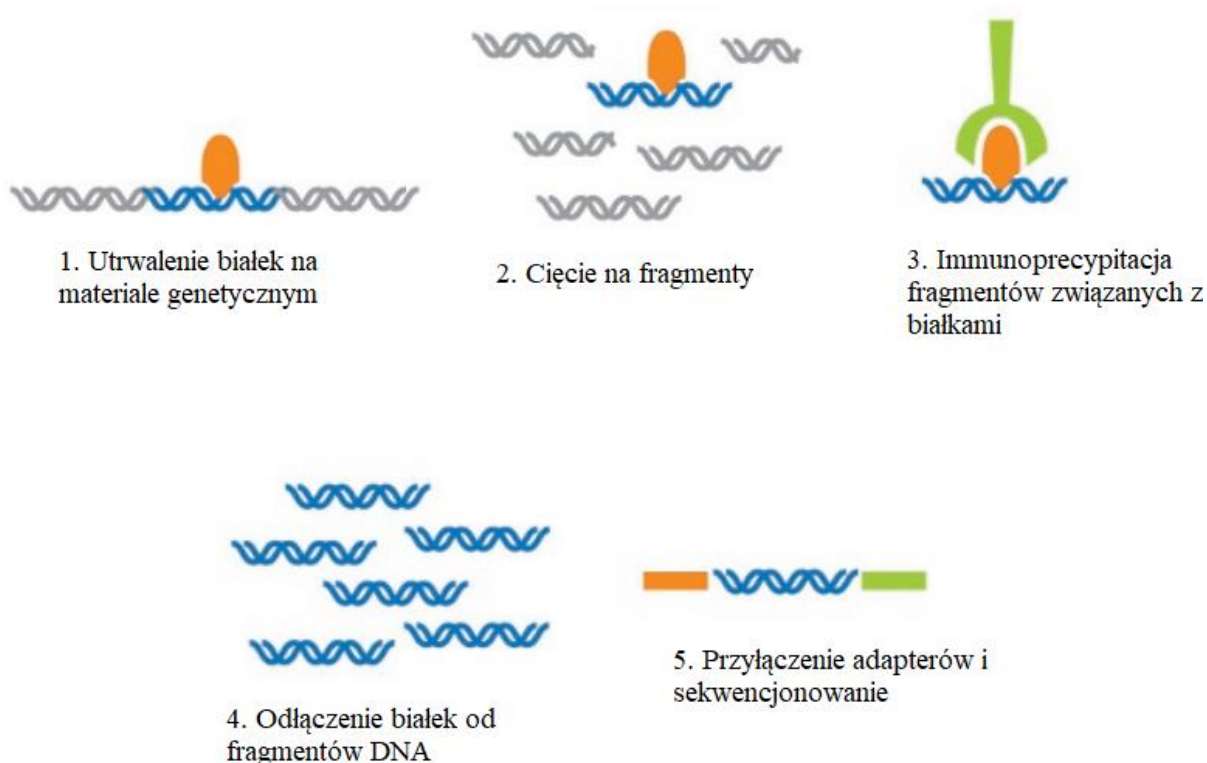
Mikromacierze wymagają znajomości genomu badanej bakterii. Technika ta jest podobna do RNA-seq, lecz jej zasada działania jest inna. Do przeprowadzenia takiego badania potrzebna jest macierz, na której w dołkach znajdują się fragmenty materiału genetycznego odpowiadające konkretnym genom bądź regionom kodującym. Z dwóch prób pobiera się RNA i tłumaczy na cDNA poprzez zastosowanie odwrotnej transkrypcji. Następnie cDNA poddaje się fragmentacji i oznacza fluorescencyjnie różnymi znacznikami [Rys. 1.6]. Nakładane na macierz próbki, z obu prób biologicznych na raz, będą hybrydować z fragmentami w dołkach macierzy, jeśli będą do nich komplementarne, a reszta próbki jest odmywana. W ten sposób z każdego dołka można odczytać sygnał fluorescencyjny, a na podstawie obliczeń ocenić jaka jest ekspresja różnicowa genów między badanymi bakteriami [30]. Jest to metoda szybsza niż RNA-seq, ale mniej czuła. Pomimo to stosuje się ją do podobnych eksperymentów jak RNA-seq. Podobnie jak w RNA-seq, wyznacza się parametry FC, jak i FDR. Technikę można stosować do porównania mutantów *Streptomyces* ze zmienionymi sekwencjami genów regulatorowych z typem dzikim, w celu zbadania znaczenia danych genów w procesach zachodzących podczas sporulacji [31], bądź do bardziej globalnego porównania zmian ekspresji w genomie w momencie hodowania bakterii w różnego rodzaju pożywkach, na przykład stałej i płynnej [32].



Rysunek 1.6 Schemat procedury zastosowania mikromacierzy.[33]

ChIP-seq stosowany jest do identyfikacji miejsc wiązania białek regulatorowych takich jak czynniki transkrypcyjne, do chromosomu. Komórki traktowane są formaldehydem w celu zatrzymania białek przy DNA, a następnie materiał genetyczny tnie się na krótkie fragmenty. Przeprowadzana jest immunoprecypitacja z przeciwciałami specyficznymi dla danego białka, wtedy oznaczone fragmenty są odzyskiwane i oczyszczane z przyłączonych białek, po czym poddawane sekwencjonowaniu [Rys. 1.7]. W ten sposób otrzymujemy zsekwencjonowane tylko fragmenty, do których przyłączone było interesujące nas białko [34]. Dzięki zastosowaniu techniki ChIP-seq możemy określić, jak zachodzą procesy związane z odczytywaniem materiału genetycznego, ale też jak wygląda

segregowanie materiału w momencie podziału komórek, czy ich rozwoju. Można określić, gdzie przyłączają się NAPy, takie jak HupS, gdy zachodzi rearanżacja materiału *Streptomyces* w momencie powstawania spor, a dokładnie podczas podziału komórek na spory [35]. Używa się też do badania regulonów czynników transkrypcyjnych w bakteriach, jak na przykład zidentyfikowanie genów kontrolowanych przez konkretne białko. Zbadanie przy użyciu ChIP-seq miejsc wiązania się konstytutywnego czynnika sigma HrdB, do DNA, pozwoliło na zgłębienie wiedzy o działaniu tego czynnika i jego roli w regulacji procesów ważnych dla komórki [36].



Rysunek 1.7 Schemat przygotowania preparatu w ChIP-seq. Zaadaptowano z [37]

Stosuje się również połączone metody NGS. Najpopularniejszym połączeniem stosowanym w badaniu regulacji genów, między innymi, u *Streptomyces* jest sparowanie RNA-seq z ChIP-seq. Umożliwia to równoczesną analizę ekspresji i lokalizacji białek regulatorowych co pozwala na dużo dokładniejszy wgląd w regulację ekspresji. RNA-seq pozwalają na wykazanie ekspresji genów na przykład po delecji, bądź nadekspresji ich regulatora, a zastosowanie po nim ChIP-seq, umożliwia identyfikację wiązań regulatora do DNA. Nałożenie wyników obu tych metod pozwala na odróżnienie genów regulowanych bezpośrednio, czyli takich do których potwierdziliśmy wiązanie się regulatora i ich zmianę ekspresji, od genów regulowanych pośrednio, gdzie wystąpiła zmiana ekspresji, lecz nie

zarejestrowano wiązania się do nich białka regulatorowego. Takie podejście pozwoliło na przykład na określenie, że Crp (*cyclic AMP receptor protein*) jest globalnym regulatorem metabolizmu. ChIP-seq wykazał wiązanie Crp do 8 z 22 klastrów genów, między innymi biosyntezy aktynorodyny, a RNA-seq po delecji genu *crp* wykazał znaczące obniżenie ekspresji genów z tych klastrów, co wskazuje na bezpośrednią regulację tych klastrów przez Crp [38].

1.4 Bazy danych i narzędzia do analizy danych sekwencjonowania

Ponieważ istnieją tysiące danych dotyczących powyżej wspomnianych zagadnień, powstały strony umożliwiające wgląd w istniejące dane pochodzące z sekwencjonowania bakterii, w tym *Streptomyces*, zawierające cały genom konkretnych gatunków, jak i wyniki z różnego rodzaju eksperymentów. Oferują one szeroką gamę możliwości wglądu w genom różnych gatunków, jak i *Streptomyces*, lecz każda z nich ma też swoje ograniczenia. Przede wszystkim, największe i najbardziej rozbudowywane bazy danych, nie będą dotyczyły konkretnie *Streptomyces*. Większość baz cieszących się popularnością to bazy ogólne, czyli nie skupiające się na konkretnym gatunku. Jako pierwszą można wymienić Gene Expression Omnibus (GEO), która jest publiczną bazą danych należącą do National Center for Biotechnology Information (NCBI) [Rys. 1.8]. GEO przechowuje i udostępnia tysiące danych pochodzących z różnych metod, takich jak RNA-seq, mikromacierze, czy ChIP-seq. W bazie GEO, zawierają się dane surowe jak i przetworzone, a wyszukiwać je, możemy po identyfikatorach eksperymentów lub w wyszukiwarce słownej. Znajdziemy tam również dane dotyczące gatunków *Streptomyces* [39]. Strona daje użytkownikowi dostęp do opisu danych, w tym, jego autorów i możliwości pobrania danych, lecz nie znajdziemy na niej żadnej możliwości obejrzenia wyników w formie graficznej. Jest to miejsce zbierające dane, a nie użytkowe narzędzie do obróbki czy prezentacji wyników.

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

GEO DataSets

Entry type: DataSets (3), Series (43), Samples (41), Platforms (0)

Organization: Customized ...

Study type: Expression profiling by array, Methylation profiling by array, Customized ...

Author: Customized ...

Attribute name: tissue (8), strain (52), Customized ...

Publication dates: 30 days, 1 year, Custom range...

[Clear all](#)
[Show additional filters](#)

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾ [Filters: Manage Filters](#)

Search results
Items: 1 to 20 of 84

<< First < Prev Page 1 of 5 Next > Last >>

1. [The TetR-like regulator SCO4385 and Crp-like regulator SCO3571 modulate heterologous production of antibiotics in *Streptomyces coelicolor* M512](#)
(Submitter supplied) Heterologous expression in well-studied model strains is a routinely applied method to investigate biosynthetic pathways. Here, we pursue a comparative approach of large-scale DNA-affinity-capturing assays coupled with semi-quantitative mass spectrometry, to identify putative regulatory proteins from *Streptomyces coelicolor* M512, that bind to the heterologously expressed biosynthetic gene clusters of the liponucleoside antibiotics caprazamycin and liposidomycin. [more...](#)
Organism: *Streptomyces coelicolor*
Type: Expression profiling by high throughput sequencing
Platform: GPL29790 4 Samples
Download data: BW
Series Accession: GSE291227 ID: 200291227

2. [Streptomyces clavuligerus Transcriptome using adpA deleted and overexpressed mutants](#)
(Submitter supplied) RNA seq analysis was performed using adpA-deleted and adpA-overexpressed *Streptomyces clavuligerus* strains to determine its regulatory effect on tunicamycin and other secondary metabolites production
Organism: *Streptomyces clavuligerus*
Type: Expression profiling by high throughput sequencing
Platform: GPL33767 4 Samples
Download data: TXT
Series Accession: GSE243267 ID: 200243267

Top Organisms [Tree](#)
Streptomyces ambofaciens ATCC 23877 (39)
Streptomyces coelicolor (15)
Streptomyces venezuelae (4)
Homo sapiens (4)
Streptomyces clavuligerus (3)
[More...](#)

Find related data
Database:

Search details
["streptomyces"[MeSH Terms] OR "Streptomyces"[Organism] OR streptomyces[All Fields]] AND rna-seq[All Fields]
 [See more...](#)

Rysunek 1.8 Wyszukiwanie dla hasła „*Streptomyces* rna-seq” w bazie danych GEO.

Podobną popularną bazą jest ArrayExpress [Rys. 1.9]. Jest to baza danych, która udostępnia dane z eksperymentów z zakresu genomiki, transkryptomiki i innych badań omicznych, w tym danych z RNA-seq, czy mikromacierzy [40]. Odnajdziemy tam szeroki zakres danych eksperymentalnych z różnych dziedzin i pozyskanych z różnych organizmów. Pozwala ona również na udostępnianie własnych danych do wglądu przez innych użytkowników, przeglądanie danych już znajdujących się w bazie danych jak i wyszukiwanie danych według różnych kryteriów, takich jak typ danych, organizm, technika, rodzaj próbki, oraz warunki eksperymentalne. Istnieją również narzędzia wyszukiwania umożliwiające łatwe odnalezienie konkretnych danych. Strona ta nie oferuje jednak wglądu do danych w sposób graficzny, są one tam dostępne do pobrania. Samodzielna analiza danych może okazać się skomplikowana dla niektórych użytkowników ze względu na różnorodność formatów.

The screenshot shows the BioStudies website interface. At the top, there is a search bar with the text 'rna-seq' and a magnifying glass icon. Below the search bar, there are navigation links: 'ArrayExpress Home', 'Browse', 'Submit', 'Help', and 'About BioStudies'. On the right, there are links for 'Feedback' and 'Login'. Below the navigation bar, there is a section for 'ArrayExpress Functional genomics data'. The main content area displays 'Search results for rna-seq'. On the left, there is a sidebar with a list of study types and their counts: 'animal - high-throughput sequencing' (392), 'cell line - high-throughput sequencing' (448), 'chip-seq' (1,062), 'methylation profiling by high throughput sequencing' (269), 'other' (448), 'rna-seq of coding rna' (11,581), 'rna-seq of coding rna from single cells' (923), 'rna-seq of non coding rna' (1,905), 'rna-seq of total rna' (382), and 'transcription profiling by array' (404). The main results area shows three search results, each with a title, date, link count, file count, and view count. The first result is 'Riboseq and Quantseq of neuroblastoma cell line SH-SY5Y expressing MAPT-AS1 constructs' (20 March 2021, 1 link, 2 files, 152 views). The second result is 'Single nucleus RNA sequencing in mouse brown adipose tissue at room temperature, cold condition and thermoneutrality' (13 July 2020, 2 links, 11 files, 1114 views). The third result is 'Whole transcriptome sequencing of splenic influenza-specific CD8+ T cells in wt and Notch1/2ko mice' (26 October 2014, 1 link, 3 files, 54 views).

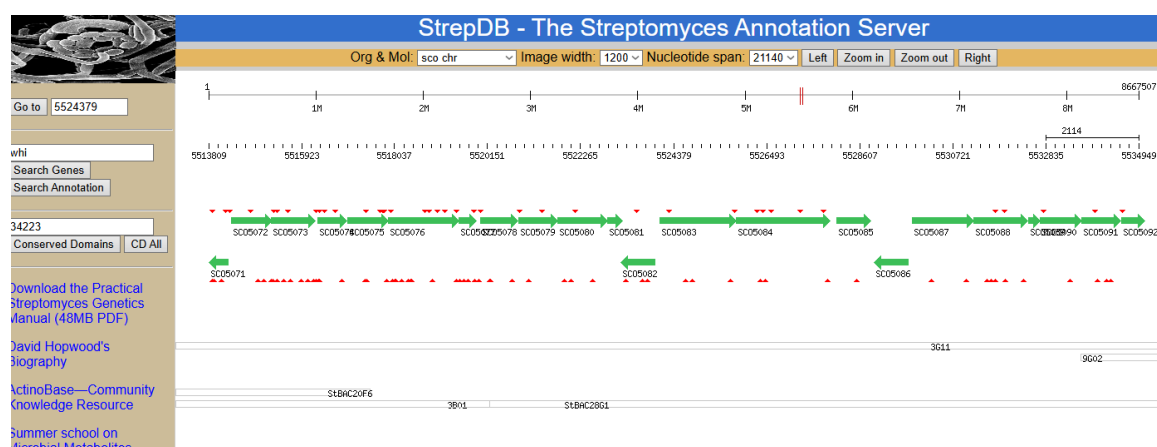
Rysunek 1.9 Wyniki wyszukiwania dla hasła „rna-seq” w bazie ArrayExpress

Istnieją również bazy, które są skoncentrowane na kolekcjonowanie danych z konkretnych gatunków, bądź chorób. Na przykład The Cancer Genome Atlas (TCGA) jest bazą danych genomowych zawierającą szczegółowe informacje o różnego rodzaju nowotworach. Projekt, który spowodował powstanie TCGA, obejmował ponad 11 tysięcy próbek nowotworowych z różnego rodzaju nowotworów i przyczynił się do powstania szerokiej gamy wyników sekwencjonowania genetycznego, i nie tylko, które wchodziły w skład tej bazy. W bazie TCGA również znajdziemy dane surowe jak i po analizach [41]. Z drugiej strony, bazą skupiającą się na konkretnych gatunkach jest iModulonDB [Rys. 1.10]. Jest ona dedykowana dla bakterii, w tym *Streptomyces*, oferuje informacje o regulacji genów i szlaków metabolicznych [42]. Głównie jest skoncentrowana na tym, jak regulacja ekspresji genów wpływa na procesy biologiczne, takie jak wzrost, produkcja metabolitów czy reakcje na stres i inne bodźce. Oferuje ona estetyczny graficzny wgląd w regulony i szlaki w niektórych bakteriiach z dokładnymi opisami danych i możliwością modyfikacji wyświetlanych wykresów. Niestety nie znajdziemy na niej informacji dotyczących *S. coelicolor* jak i *S. venezuelae*, w naszym przypadku, gatunków nas interesujących.



Rysunek 1.10 Prezentacja wyników dla WhiH *Streptomyces albidoflavus* w bazie iModulonDB

Bazą najbardziej interesującą w temacie samego *Streptomyces* jest StrepDB. Oferuje ona szeroki zakres zasobów związanych z badaniami nad *Streptomyces*, [Rys. 1.11] jest to baza zawierająca genomy różnych gatunków, informacje o BGC metabolitów i inne podobne dane transkryptomiczne. Strona udostępnia narzędzia do analizy danych takie jak BLAST, jak i dokładne opisy genów znajdujących się na wykresach i domen. Natomiast na StrepDB nie dostaniemy możliwości obejrzenia najnowszych danych wraz poziomami ekspresji, a tym bardziej nie znajdziemy na niej opcji obejrzenia własnych danych. Służy raczej ona do dokładnego wglądu w istniejące już, pełne, genomy gatunków *Streptomyces*.



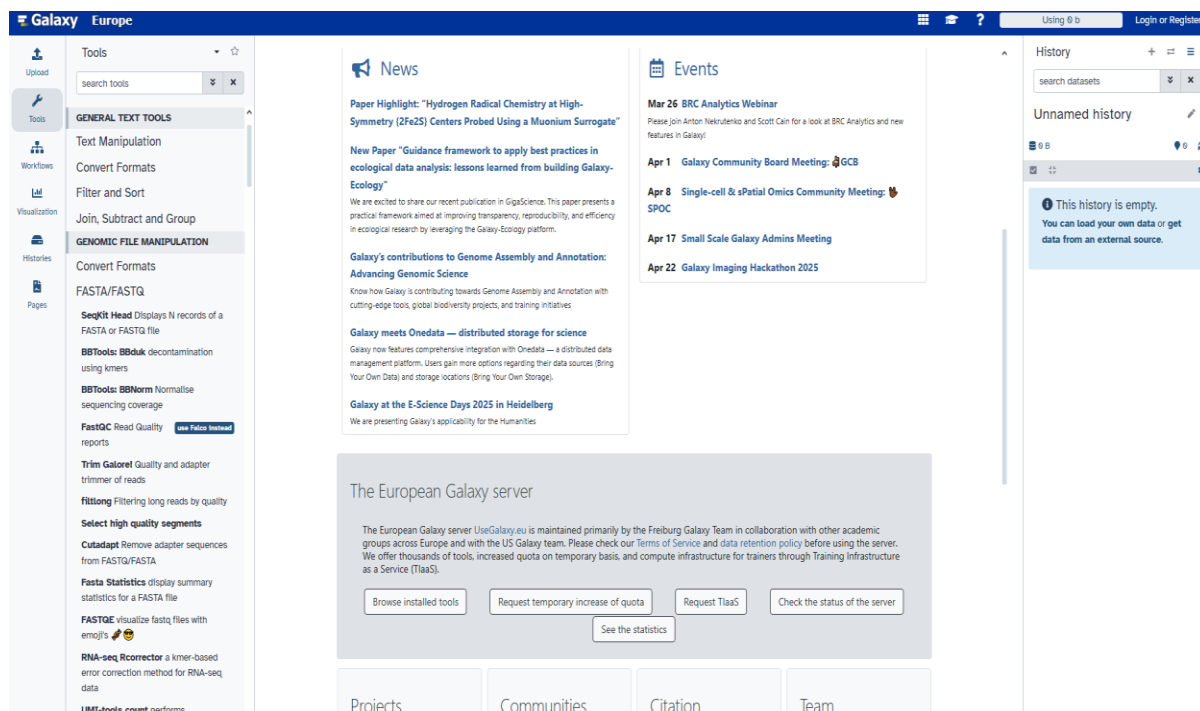
Rysunek 1.11 Klaster biosyntezy aktynorodyny wyświetlony w bazie StrepDB

Same bazy danych są nie wystarczające w analizie wyników dotyczących sekwencjonowania genetycznego czy innego typu eksperymentów. Istnieje wiele programów i stron, służących jako narzędzia do analiz danych pozyskanych samemu, czy też zaczerpniętych z wyżej wymienionych baz danych. Dzięki zastosowaniu takich narzędzi możemy przeanalizować i zwizualizować nasze dane w sposób graficzny, co zdecydowanie ułatwia ich zrozumienie. W przypadku danych z sekwencjonowania, narzędzia te posiadają specjalne funkcje przystosowane właśnie do takiego typu wyników, takie jak usuwanie adapterów z odczytanych sekwencji, czy mapowanie odczytów do genomu.

Jedną z częściej używanych platform dostępnych za darmo jest Galaxy [Rys. 1.12]. Jest to platforma bioinformatyczna pozwalająca na analizę danych biologicznych, a jej działanie i to co oferuje, jest specjalnie zaprojektowane dla naukowców, którzy nie mają doświadczenia w programowaniu. Galaxy jest dostępne przez przeglądarkę internetową. Platforma Galaxy zawiera dziesiątki różnego rodzaju narzędzi do analizy bioinformatycznej, między innymi do analizy sekwencjonowania, takich jak FastQC (ocena jakości surowych odczytów), Trimmomatic (usuwanie adapterów), czy edgeR (analiza różnicowej ekspresji). Galaxy umożliwia użytkownikowi wygodę pracy dzięki opcjom dostępnym na platformie. Możemy zapisać i ponownie uruchomić przepływ pracy (workflow), nad którym pracujemy, a także współpracować z innymi użytkownikami. Każdy krok wykonany na Galaxy jest rejestrowany i może być odtworzony, co zapewnia reprodukowalność badań. W gamie narzędzi dostępnych na Galaxy, znajdziemy również takie, które pozwalają na prezentację wyników w formie graficznej [43]. Należy jednak podkreślić, że jest to narzędzie bioinformatyczne, więc musimy mieć dane gotowe do analizy, bo na stronie Galaxy nie znajdziemy dostępnych danych.

Istotne również w analizach sekwencjonowania na przykład RNA-seq, jest mapowanie odczytów do genomu. Bez tego procesu nie byłibyśmy w stanie określić, jak wyglądała ekspresja poszczególnych genów. Przykładem narzędzia do mapowania odczytów jest bowtie2 lub hisat dostępne na platformie Galaxy. Po uzyskaniu wyniku mapowania najczęściej przeprowadza się jego analizę, zwykle w formie graficznej. Do tego zadania posłużyć może Integrative Genomics Viewer (IGV) [44]. Jest to interaktywna przeglądarka genomowa, która pozwala na wizualizację mapowania genomu. Możemy na IGV oglądać dane, które mogą zostać załadowane w różnego rodzaju formatach takich BAM, czy BED, a do tego korzystać z obecnych już na IGV genomów referencyjnych lub załadować własny w formacie FASTA, co pozwoli na analizę odczytów u interesującego nas gatunku. Jest to

program łatwy w użyciu i dostępny za darmo w wersji przeglądarkowej jak i desktopowej [45].



Rysunek 1.12 Interfejs platformy Galaxy z rozwiniętymi narzędziami do modyfikacji plików genomowych.

Podobnym narzędziem do Galaxy jest iDEP. Jest również stroną, która funkcjonuje raczej jako narzędzie bioinformatyczne aniżeli baza danych, a do tego kładzie nacisk głównie na przedstawianie graficzne wyników [46]. Na iDEP jesteśmy w stanie przeanalizować dane z RNA-seq i mikromacierzy, a następnie wyświetlić przeanalizowane dane w formie wykresów i tabel, które są modyfikowalne. Jest to narzędzie dostępne online bez potrzeby pobierania oprogramowania i umożliwia użytkownikowi analizę różnicowej ekspresji genów oraz powiązanych szlaków metabolicznych i funkcjonalnych. Niestety, jak już zostało wspomniane, iDEP nie jest bazą danych, co oznacza, że poza plikiem „demo” nie znajdziemy tam innych danych, które będziemy mogli obejrzeć, jedynie te, które sami załadujemy na stronę. Nie jest to jedyna tego typu platforma do szybkich analiz i wizualizacji, z podobnych można wymienić GenePattern, ToppGene czy ShinyGO [47], [48], [49].

2. Cel pracy

Celem pracy jest stworzenie narzędzia, a zarazem bazy danych, które poprzez połączenie zalet istniejących już programów, ułatwiłoby analizę i generowanie wykresów porównawczych dla danych transkryptomicznych *S. coelicolor* i *S. venezuelae*. Zapotrzebowanie na taką aplikację wynika z braku szybkiego dostępu do uporządkowanych tego typu danych z tych dwóch gatunków, na istniejących już platformach. Program miałby nie tylko gromadzić dane, ale również umożliwiać ich analizę, porównanie oraz wizualizację różnic w ekspresji. Jedną z funkcji byłaby możliwość dodawania nowych danych oraz wgrywania własnych wyników. Aplikacja miałaby zapewniać interaktywną nawigację po genomie, filtrację danych oraz generowanie wykresów. Dodatkowo przewidziano opcję pobierania wyników i dokumentację użytkownika.

Cele pracy:

- Gromadzenie danych z RNA-seq, mikromacierzy i ChIP-seq dla *S. coelicolor* i *S. venezuelae*.
- Udostępnianie i wizualizacja danych z możliwością analizy różnic w ekspresji.
- Możliwość wgrywania własnych danych i ich porównania z bazą.
- Opcja filtrowania wyników według określonych kryteriów.
- Porównywanie różnych zbiorów danych w celu wykrywania podobieństw i różnic.
- Możliwość pobrania wizualizacji wyników do publikacji.
- Stałe rozszerzanie bazy o nowe dane nawet po zakończeniu projektu.

3. Materiały i metody

Dane, które zostały zgromadzone z publikacji, wybierane były na podstawie opisywanych w nich eksperymentów i interesujących nas gatunków. Wszystkie dane załączone w aplikacji pochodzą z eksperymentów dotyczących *S. coelicolor* i *S. venezuelae*, i wykonywanych na nich badań z zakresu NGS lub mikromacierzy [Tab. 3.5]. Oczywiście dane te są dostępne dla każdego, lecz problem techniczny pojawia się w momencie, gdy próbujemy porównywać ze sobą wyniki zaprezentowane w artykułach. Często autorzy skupiają się tylko na konkretnych genach interesujących ich w swoich badaniach, dlatego w samym tekście publikacji trudno o szczegółowy wynik z całego sekwencjonowania. Bardziej rozbudowane dane można znaleźć w suplemencie znajdujących publikacji. Spotkać można się z wynikami w postaci tabeli w arkuszu kalkulacyjnym Excel, tabeli zapisanej w formacie PDF, czy też tabel w programie Word. Autorzy rzadko udostępniają dane w postaci czystej liczby zliczeń odczytów przypadających na poszczególne geny. Wszystkie dane, które znajdują się w aplikacji są przenalizowane i znormalizowane w różny sposób. Wartości liczbowe pochodzące z danych nie zostały w żaden sposób zmienione przed dodaniem do programu, a tylko przekształcone w celu prezentacji wartości udostępnionych przez oryginalnych autorów publikacji. Oprócz tego większość wyników z publikacji nie zawiera danych z całego sekwencjonowania, a przedstawia jedynie logFC i FDR dla genów uznanych za istotne statystycznie przez badaczy.

Program powstał przy użyciu języka programowania R [50], a dokładniej graficznego środowiska RStudio [51]. R to język programowania i środowisko obliczeniowe głównie do analizy statystycznej, wizualizacji danych oraz obliczeń naukowych. Jest on szeroko stosowany w bioinformatyce. Pozwala on na obsługę dużych zbiorów danych, dzięki czemu idealnie nadaje się do pracy z danymi eksperymentalnymi, jak w tym przypadku, danych z sekwencjonowania. Pozwala on na tworzenie zaawansowanych wykresów i wizualizacji danych, szczególnie dzięki pakietom, czyli dodatkowym zestawom funkcji, do niego dostępnych, takich jak ggplot2 [52]. Język R jest też oprogramowaniem open-source i ma ogromną społeczność rozwijającą dodatkowe pakiety, a ze względu na zainteresowanie ze strony dziedziny bioinformatyki, powstają takie, które są dedykowane do danych biologicznych. W procesie tworzenia aplikacji zastosowane zostały różnorodne pakiety [Tab. 3.1], na przykład gggenes [53], czyli rozszerzenie ggplot2, umożliwiające tworzenie wykresów strzałkowych obrazujących geny. Natomiast RStudio to zintegrowane środowisko programistyczne dla R, które ułatwia pracę w tym języku. Pozwala ono na lepszą organizację

kodu i danych, autouzupełnianie kodu, debugowanie i podgląd danych czy zarządzanie pakietami. Aplikacja powstała przy użyciu pakietu Shiny [54]. Jest to pakiet, który powstał, by pozwolić na tworzenie interaktywnych aplikacji webowych, czyli takich, które uruchamia się w przeglądarce. Shiny ułatwia tworzenie aplikacji bez potrzeby znajomości HTML czy JavaScript. Aplikacje powstałe przy użyciu tego pakietu pozwalają na wgląd do udostępnionych danych i interakcję z nimi, w tym ich wizualizację.

Tabela 3.1 Zbiór pakietów załadowanych w programie RStudio, które zostały użyte do stworzenia aplikacji.

Pakiety użyte w RStudio	
Nazwa pakietu	Funkcja
shiny	Umożliwia stworzenie aplikacji webowych [54]
shinythemes	Dodaje gotowe motywy do aplikacji Shiny [55]
ggplot2	Tworzenie wykresów [52]
gggenes	Rozszerzenie ggplot2, służące do wizualizacji struktur genowych [53]
dplyr	Rozszerza funkcje służące do manipulacji danymi [56]
tidyr	Rozszerza funkcje do przekształcania danych [57]
patchwork	Umożliwia łączenie wielu wykresów ggplot2 w jeden układ [58]
tidyHeatmap	Tworzenie heatmap w stylu tidyverse [59]
BiocManager	Umożliwia instalacje i zarządzanie pakietami Bioconductor [60]
stringr	Służy do łatwej i spójnej pracy z tekstem (łańcuchami znaków).[61]
bslib	Stylizowanie aplikacji w Shiny [62], [63]
ggvenn	Rozszerzenie ggplot2, umożliwia rysowanie wykresów Venn'a [63]
ggupset	Rozszerzenie ggplot2, umożliwia rysowanie wykresów upset [64]
pdftools	Umożliwia czytanie i przetwarzanie plików PDF w R [65]
png	Obsługa plików graficznych PNG [66]

Tak jak człowiek nie jest w stanie łatwo poradzić sobie z wieloma plikami w różnych formatach, podobnie jest z programem komputerowym. Aby aplikacja w Shiny w R działała, dane muszą być ujednolicone w jednym formacie, najlepiej zgodnymi z zasadami „tidy data” [67]. Dane „tidy” charakteryzują się głównie trzema zasadami; każda kolumna to jedna zmienna (np. temperatura, czas, kontrast), każdy wiersz to jedna obserwacja (np. gen, chromosom), każda komórka zawiera jedną wartość. Niestety zdecydowana większość danych, które można znaleźć w suplementach publikacji czy bazach danych nie jest zapisane w formacie tidy. Dlatego należało przeformatować wszystkie zebrane dane na „tidy”, a aby to osiągnąć przygotowano ujednolicony układ/format danych do aplikacji zawierający wszystkie najważniejsze dla nas wartości. Dane dla RNA-seq, wyjściowo zawierają identyfikatory genów, wartość logFC, FDR i kolumnę nazwaną „add_variable” od „additional variable” czyli dodatkowej zmiennej, w którą można wpisać dowolną nazwę porównania czy zestawienia między godzinami pobrania próbki [Tab. 3.2]. Każde dane, które trafiły do programu, bądź będą dodane przez użytkownika są przyrównywane do genomu referencyjnego w celu dodania lokalizacji danych genów na genomie. Format danych został ujednolicony ręcznie poprzez modyfikację plików wyjściowych pochodzących ze zgromadzonej bazy danych [Tab. 3.5]. Dokładny schemat transformacji danych z wyjściowych na te już dostępne w aplikacji, różnił się dla poszczególnych przypadków, ze względu na wysokie zróżnicowanie formatu danych wyjściowych. Dane są przeformatowane, bez ingerencji w wyniki w nich zawarte, zmienione jedynie jest ułożenie wyników wyjściowych. Dla użytkownika, dostępna w aplikacji jest opcja pobrania każdego z szablonów danych w formacie „xlsx”.

Tabela 3.2 Szablon formatu danych RNA-seq w aplikacji.

gene	logFC	FDR	add_variable
SCO0001	0	1	variable_1
SCO0005	0	1	variable_1
SCO0003	0	1	variable_1
SCO0003	0	1	variable_2
SCO0005	0	1	variable_2
SCO0001	0	1	variable_2
SCO0004	0	1	variable_2
SCO0003	0	1	variable_3

W przypadku danych pochodzących z eksperymentów ChIP-seq szablon domyślny dla aplikacji nieco się różni. Zawiera on wyjściowo również cztery kolumny, lecz nie jest przyrównywany do genomu, ponieważ zawiera on lokalizacje miejsc wiązań [Tab. 3.3]. W kolumnie „name” znajdują się warunki badane w eksperymencie.

Tabela 3.3 Szablon formatu danych ChIP-seq w aplikacji.

chrom	chromStart	chromEnd	name
chromosome1	1	100	variable_1
chromosome1	101	200	variable_1
chromosome1	201	300	variable_1
chromosome1	301	400	variable_1
chromosome1	1	100	variable_2
chromosome1	101	200	variable_2
chromosome1	201	300	variable_2
chromosome1	301	400	variable_2

Istnieje również dodatkowa opcja wczytywania danych RNA-seq omijająca przekształcanie ręcznie w format „tidy”. Przygotowany został szablon danych łatwy do przekształcenia w Excelu. Składa się on z trzech części. Pierwsza kolumna musi zostać nazwana „gene” i znajdują się tam wszystkie identyfikatory genów. Następne dwie to część logFC i FDR. Zawierać ona będzie tyle kolumn, ile użytkownik posiada warunków, czy zmiennych w eksperymencie. Aby dane zostały poprawnie przekształcone przez program do formatu „tidy”, należy nazwać kolumny nazwą warunku i nazwą odpowiedniego rodzaju wartości danych. Na przykład, jeśli dokonane zostały pomiary hodowli mutantu i typu dzikiego w 48, 58 i 64 godzinie i zostało z porównania dwóch szczepów obliczone logFC, kolejne kolumny należy nazwać „48hlogFC”, „58hlogFC” i „64hlogFC”. Odpowiednio odpowiadające im kolumny FDR, należy nazwać „48hFDR”, „58hFDR” i „64hFDR” jak na przykładzie [Tab. 3.4]. Program przekształci dane do formatu „tidy”, gdzie geny wystąpią 3-krotnie, w nowo-powstałej kolumnie logFC będą znajdowały się wszystkie wyniki logFC, a w kolumnie FDR, wszystkie wyniki FDR, natomiast przedrostki „48h”, „58h” i „64h” znajdą się w czwartej kolumnie nazwanej „add_variable”. Polska wersja programu Excel jako

znak interpunkcyjny w liczbach niecałkowitych używa przecinka, automatycznie przecinki w kolumnach przekształcane są na kropki, a zawartości kolumn nadawany jest typ wartości numerycznych, ponieważ język programowania R, jako znaku interpunkcyjnego w takich liczbach używa kropek.

Tabela 3.4 Szablon formatu danych RNA-seq do obróbki przez program na format „tidy”.

gene	48hlogFC	58hlogFC	64hlogFC	48hFDR	58hFDR	64hFDR
SCO0033	0,12	0,12	0,12	2,00E-01	2,00E-01	2,00E-01
SCO0044	0,59	0,59	0,9	1,00E-02	1,00E-02	1,00E-02
SCO0055	0,62	0,62	0,62	3,00E-02	3,00E-02	3,00E-02
SCO0058	0,28	0,28	0,28	3,00E-09	3,00E-09	3,00E-09
SCO0064	0,45	0,45	0,45	2,00E-04	2,00E-04	2,00E-04

Tabela 3.5. Zbiór publikacji, z których pochodzą dane wykorzystane w bazie danych aplikacji.

Dane dostępne w aplikacji				
Opis publikacji	W badaniu porównano	Typ eksperymentu	Gatunek	Nazwa w aplikacji
Rola NAP w regulacji ekspresji genów i odpowiedzi na stres [18].	Mutanty z delecją <i>hupA</i> i/lub <i>hupS</i> między sobą i WT między sobą w różnych warunkach	RNA-seq	<i>S. coelicolor</i>	hupAS_Strzalka_2024
Zbadanie globalnej odpowiedzi transkrypcyjnej na zmiany chromosomu wywołane szybką relaksacją DNA [68].	Mutanty ze zwiększoną ekspresją TopA z WT	RNA-seq	<i>S. coelicolor</i>	TopA_Szafran_2019
Zbadanie systemu dwuskładnikowego SatKR w regulacji transkrypcji klastra genów wrażliwych na superskręcenie DNA [69].	Mutanty z delecją genów <i>satkR</i> i <i>sitkR</i> oraz mutanty z obniżoną ekspresją TopA	RNA-seq	<i>S. coelicolor</i>	SatKR_Gongerowska_2021

Znaczenie systemu AbrB1/B2 w produkcji antybiotyków i odporności antybiotykowej [70].	Mutant z delecją <i>abrB</i> z WT	RNA-seq	<i>S. coelicolor</i>	AbrB1_Nieta_2020,
Znaczenie AbrC3 w regulacji produkcji antybiotyków [71].	Mutant z delecją <i>abrC3</i> z linią dziką rodzicielską	Mikromacierz	<i>S. coelicolor</i>	AbrC3_Rico_2014
Znaczenie Aor1 w regulacji biosyntezy antybiotyków [27].	Mutant z delecją <i>aor1</i> z WT	RNA-seq	<i>S. coelicolor</i>	Aor1_Antoraz_2017
Znaczenie, rodzaj regulacji i funkcje ArgR w hodowli płynnej. [72]	Mutant z delecją <i>argR</i> z WT	RNA-seq	<i>S. coelicolor</i>	argr_Botas_2018
Znaczenie BldD w rozwoju morfologicznym i produkcji antybiotyków w czasie wzrostu wegetatywnego [73].	Mutant z delecją <i>bldD</i> z WT	Mikromacierz	<i>S. coelicolor</i>	BldD_denHengst_2010

Wpływ BldC na przebieg cyklu życiowego [74].	Mutant z delecją <i>bldC</i> z WT	RNA-seq	<i>S. venezuelae</i>	BldC_Bush_2018_sven
Rola systemu DraR-K w różnicowaniu morfologicznym i fizjologicznym [75].	Mutant z delecją <i>draR-K</i> z linią dziką rodzicielską	Mikromacierz	<i>S. coelicolor</i>	Drar-K_Yu_2014
Scharakteryzowanie funkcji, mechanizmów regulacyjnych i roli fizjologicznej nowej grupy czynników sigma ECF42 [76].	Różne mutanty delecyjne lub z nadekspresją ECF42	RNA-seq	<i>S. venezuelae</i>	ECF42_Liu_2018_sven
Znaczenie GlnR w regulacji odpowiedzi na niedobór azotu [77].	Mutant z delecją <i>glnR</i> z WT	Mikromacierz	<i>S. venezuelae</i>	GlnR_Pullan_2011
Wpływ OhkA na regulacje metabolizmu drugorzędowego i różnicowanie morfologiczne.[28]	Mutant z delecją <i>ohkA</i> z WT	RNA-seq	<i>S. coelicolor</i>	OhkA_Lu_2011

Znaczenie OsdR jako regulatora odpowiedzi na stres oksydacyjny [78].	Mutant z delecją <i>osdR</i> z WT	Mikromacierz	<i>S. coelicolor</i>	OsdR_Urem_2016
Rola czynnika sigma R w kontroli jakości białek w stresie disiarczowym [79].	Mutant z delecją <i>sigR</i> z WT	RNA-seq	<i>S. coelicolor</i>	sigR_Kallifidas_2010
Zbadanie celów dla SoxR [80].	Mutant z delecją <i>soxR</i> z WT	RNA-seq	<i>S. coelicolor</i>	soxR_Naseer_2014
Identyfikacja nowych genów biorących udział w sporulacji. [31]	Mutant z delecją <i>whiA</i> lub <i>whiH</i> z WT	Mikromacierz	<i>S. coelicolor</i>	whiAH_Salerno_2013
Analiza transkryptomyczna różnicowania w stałych hodowlach sporulujących [32].	Hodowle płynną z hodowlą stałą	Mikromacierz	<i>S. coelicolor</i>	SolidLiquidDiff_Yague_2013
Szczegółowe zbadanie procesów transkrypcyjnych i translacyjnych [29].	Różne etapy cyklu życiowego	RNA-seq	<i>S. coelicolor</i>	growth_phases_Yeong_2016

Wpływ szoku etanolowego na transkrypcję i metabolizm komórek [81].	Komórki pod wpływem szoku etanolowego z WT	RNA-seq	<i>S. venezuelae</i>	ETOHshock_Sekurova_2022_sven
Globalny wpływ BldB na regulację ekspresji komórek.[82]	Mutant z delecją <i>bldB</i> z WT	RNA-seq	<i>S. venezuelae</i>	BldB_Avramova_2023_sven
Odpowiedź regulatora LexA na stres wywołany degradacją DNA przy użyciu mitomicyny C (MMC) [83]	Komórki poddane MMC z WT	RNA-seq	<i>S. venezuelae</i>	MMC_Stratton_2022_sven
Oddziaływanie Ccr1 na metabolizm drugorzędowy bakterii [84]	Mutant z delecją <i>ccr1</i> z WT	RNA-seq	<i>S. coelicolor</i>	ccr1_Lei_2024
Rola Gbn (GATC-binding NAP) w rozwoju <i>Streptomyces</i> [85]	Mutanty z delecją <i>gbn</i> z WT	RNA-seq	<i>S. coelicolor</i>	Gbn_Du_2022
Rola Gbn (GATC-binding NAP) w rozwoju <i>Streptomyces</i> [85]	---	ChIP-seq	<i>S. coelicolor</i>	Gbn_chipseq_Du_2022

Analiza regulonu HrdB w fazie wzrostu wegetatywnego [36].	---	ChIP-seq	<i>S. coelicolor</i>	HrdB_chipseq_Smidova_2018
Zbadanie współdziałania TopA i HupA w utrzymaniu organizacji chromosomu podczas wzrostu [86].	---	ChIP-seq	<i>S. coelicolor</i>	hupAS_chipseq_Strzalka_2024
Analiza rearanżacji przestrzennej chromosomu podczas powstawania spor [35].	---	ChIP-seq	<i>S. venezuelae</i>	SMChupS_Szafran_2021

4. Wyniki

Aplikacja została przygotowana jako narzędzie bioinformatyczne, które ma na celu pomoc badaczom skupiającym się na transkryptomicznych aspektach badań *S. coelicolor* i *S. venezuelae*. Ważne jest, aby użytkownik, który będzie chciał skorzystać z aplikacji, był w stanie zrozumieć co robią jej poszczególne funkcje i przyciski w niej dostępne. W pierwszym rozdziale wyników szczegółowo opisano jak prawidłowo korzystać z aplikacji. Umożliwi to użytkownikowi efektywne wykorzystanie aplikacji w analizie, jak i poznanie wszystkich opcji dostępnych w programie. Drugi rozdział sekcji wyniki, przedstawia działanie aplikacji opisane na podstawie kilku przykładów analiz danych eksperymentalnych. Postawiony został nacisk, nie na zaprezentowanie, każdej możliwości, którą użytkownik może chcieć przeanalizować, a raczej kilku konkretnych przypadków, korzystających z jak największej ilości dostępnych funkcji. Poszczególne przykłady użycia aplikacji, przedstawiają wszechstronne zastosowanie programu, do analizy danych w nim dostępnych, jak i danych własnych, załadowanych przez użytkownika, podczas zestawienia ze sobą różnego rodzaju eksperymentów transkryptomicznych.

4.1 Program Strep.N.A

Powstała aplikacja nosi nazwę Strep.N.A (*Streptomyces* NGS Analysis). Program można pobrać z GitHub autora i uruchomić go lokalnie, po zainstalowaniu RStudio i pakietów opisanych w tabeli [Tab. 3.1], lub skorzystać z wersji dostępnej online na serwerze.

Link do serwera:

http://microbesinwroclaw.biotech.uni.wroc.pl:3838/zmm_apps/Praca_magisterska_Marek_Skrzynski/

Link do github:

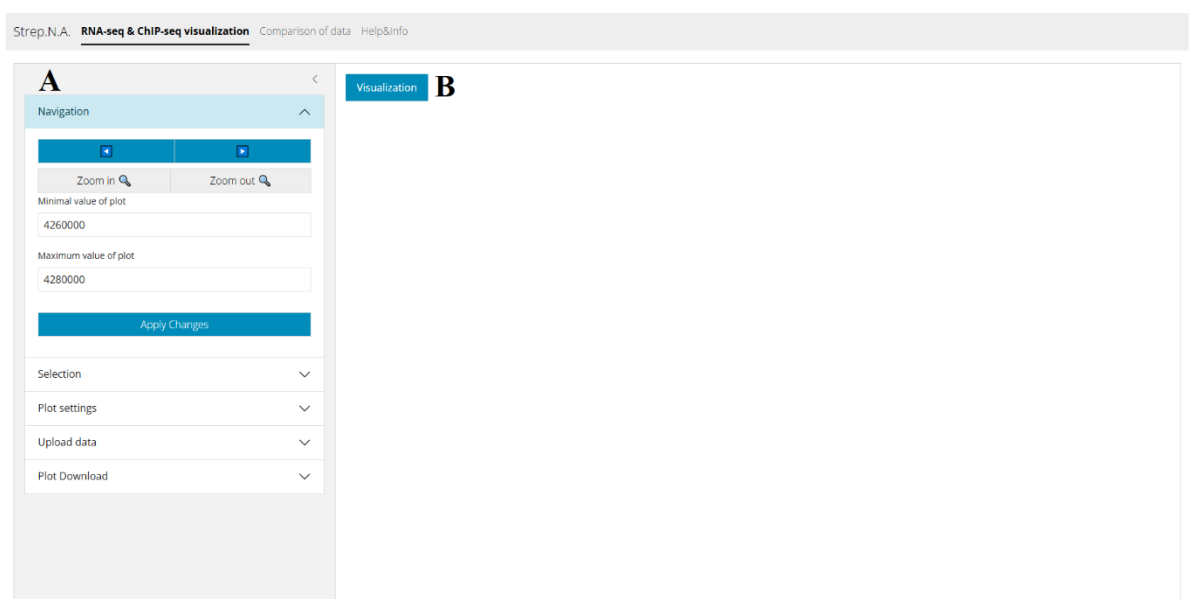
https://github.com/Majusz1320/Praca_magisterska_Marek_Skrzynski

Strep.N.A posiada trzy główne strony: „RNA-seq & ChIP-seq visualization”, „Comparison of data” i „Help&Info”. Dwie pierwsze z wymienionych są stronami funkcjonalnymi, to znaczy, możemy na nich ładować dane czy wyświetlać je w celu analiz. Strona „Help&Info” jest stroną przeznaczoną jako pomoc dla użytkownika w nawigacji na stronie, jak i zawiera niezbędne informacje dotyczące artykułów składających się na bazę danych. Strony funkcjonalne, podzielić możemy na dwie części: panel boczny, zawierający

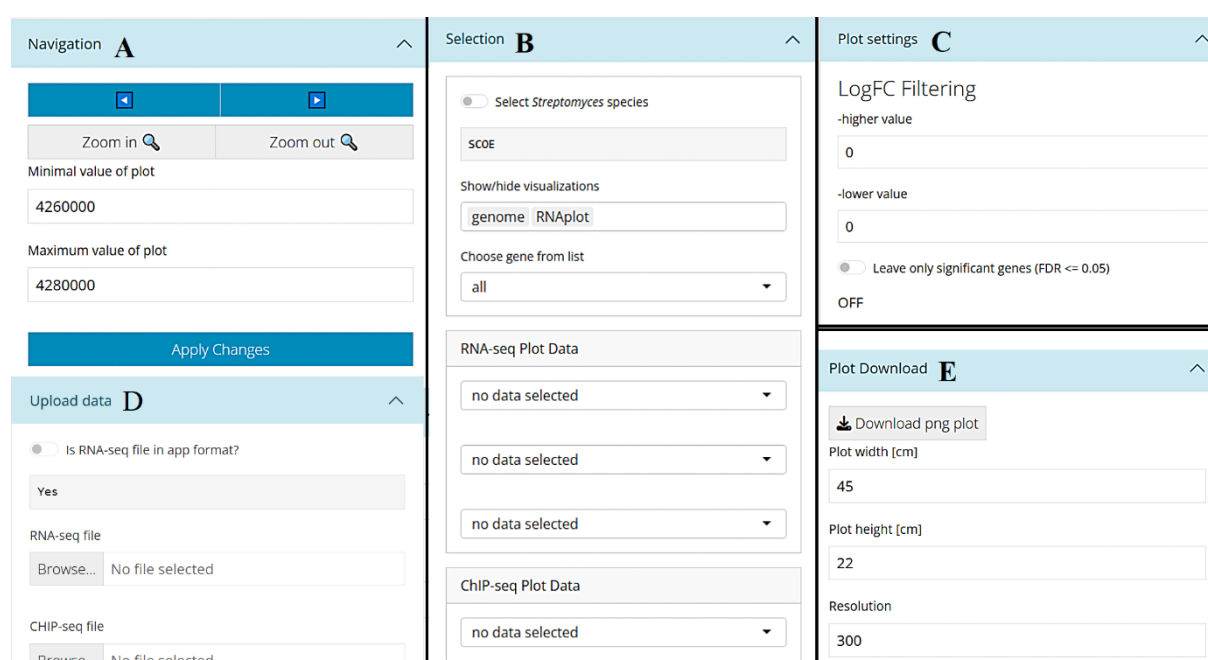
wszystkie przyciski umożliwiające ładowanie danych czy nawigację po wyświetlanych danych i panel główny, na którym znajdować się będą wszystkie wyświetlane wykresy oraz tabele. Aby ułatwić zrozumienie na co pozwala Strep.N.A potrzebne jest omówienie wszystkich elementów składających się na program.

4.1.1 Podstawowe informacje po włączeniu

Po uruchomieniu programu znajdujemy się domyślnie na stronie „RNA-seq & ChIP-seq visualization” [Rys. 4.1]. Aplikacja jest gotowa do działania od razu po włączeniu. Po lewej stronie ekranu widzimy panel boczny zawierający wszystkie przyciski potrzebne do nawigacji i użytkowania tej strony [Rys. 4.1 (A)], natomiast prawą stronę ekranu zajmuje panel główny, na tej stronie, zawierający jedną podstronę „Visualization” [Rys 4.1 (B)]. Na panelu głównym domyślnie nic nie jest wyświetlane. Panel boczny zawiera na tej stronie pięć zakładek, z czego tylko pierwsza, „Navigation”, domyślnie jest rozwinięta [Rys. 4.2 (A)]. Na tej zakładce znajdują się przyciski służące do nawigacji po wyświetlonym wykresie. Strzałki w prawo i lewo umożliwiają użytkownikowi przesuwanie się po osi x wyświetlanego wykresu o wartość dziesięciu tysięcy par zasad, to znaczy, od granicznych wartości wykresu w przypadku naciśnięcia na przykład przycisku w prawo, wartości te zostaną zwiększone o dziesięć tysięcy. Przyciski „Zoom in” i „Zoom out” działają na podobnej zasadzie, lecz pozwalają na zwiększanie różnicy między wartościami granicznymi wykresu lub jej zmniejszanie co powoduje, „przybliżenie” bądź „oddalenie” widocznego wykresu. Poprzez zmienianie wartości „Minimal-” i „Maximum value of plot” możemy wybrać dokładne miejsce do obserwacji. Wartości te można zmieniać wpisując wybraną przez siebie liczbę lub naciskając strzałki znajdujące się obok wartości widoczne po najechaniu myszką na panel tekstowy zawierający aktualnie ustawioną wartość liczbową. Używanie przycisków nawigacji prawo, lewo, „zoom in” i „zoom out” zachodzi automatycznie, natomiast manualną zmianę wartości granic wyświetlanego wykresu musimy zatwierdzić przyciskiem znajdującym się u dołu zakładki „Navigation”, nazwanego „Apply Changes”.



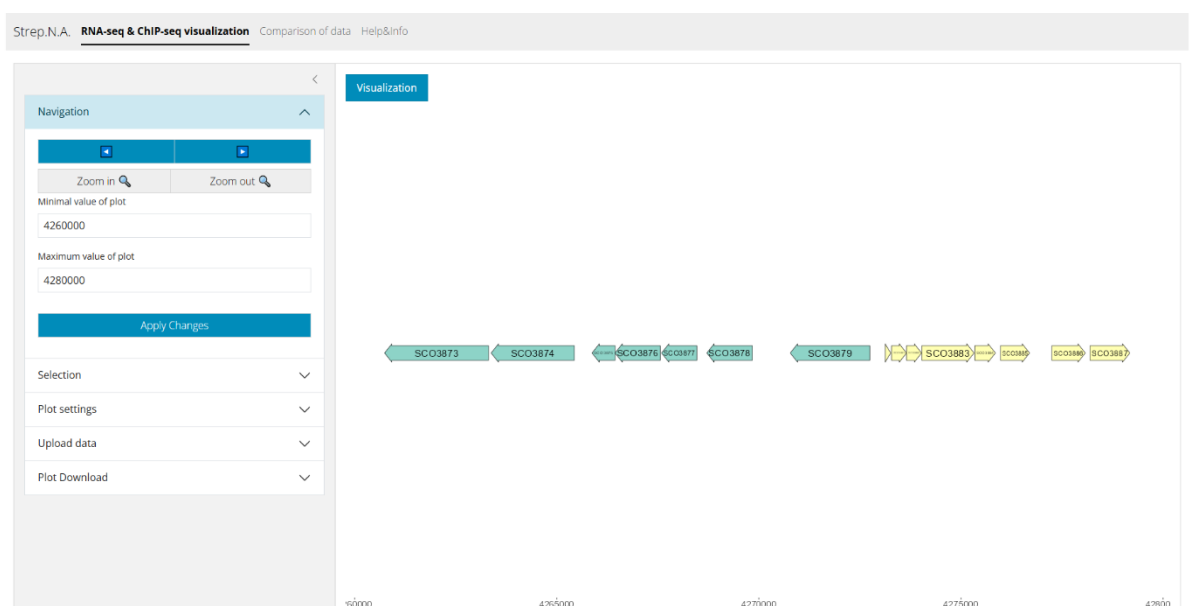
Rysunek 4.1 Główny panel aplikacji Strep.N.A widniejący na ekranie po jej włączeniu. U góry strony obok nazwy programu, zaznaczona jest strona, na której się znajdujemy, „RNA-seq & ChIP-seq visualization”. Widoczny jest boczny panel przycisków (A) i panel główny, na którym wyświetlać się będą wykresy (B).



Rysunek 4.2 Panel boczny strony „RNA-seq & ChIP-seq visualization” z rozwiniętymi (obok siebie) w orientacji poziomej zakładkami „Navigation” (A), „Selection” (B), „Plot settings” (C), „Upload data” (D), „Plot download” (E).

4.1.2 Panel boczny i funkcje „RNA-seq & ChIP-seq visualization”

Pierwszą rzeczą, którą należy zrobić po włączeniu aplikacji to naciśnięcie przycisku „Apply changes”. Bez tego żadne inne interakcje nie będą działały w programie. Naciśnięcie przycisku „Apply Changes” spowoduje rozpoczęcie reagowania programu na dalsze zmiany, a także na panelu głównym wyświetlony zostanie wykres strzałkowy genomu wybranego domyślnie gatunku (*S. coelicolor*) [Rys. 4.3]. Gdy widoczny jest wykres genomowy, możemy przejść do wyboru, co ma zostać wyświetlone na wykresie. Należy zatem wrócić do panelu bocznego, a dokładnie do zakładki „Selection” [Rys. 4.2 (B)] Zakładka ta jest podzielona na 3 części. Pierwszą są ustawienia ogólne wyświetlanych wykresów. Jako pierwszy widnieje przełącznik opisany „Select *Streptomyces* species”, służący do wyboru gatunku *Streptomyces*, na którym zamierzamy pracować. Domyślnie gatunkiem wybranym jest *S. coelicolor* stąd pod przełącznikiem w panelu tekstowym wyświetlane jest „SCOE” czyli skrótowy zapis tego gatunku (*S*-treptomyces *COE*-licolor). Naciśnięcie przełącznika spowoduje zmianę wyświetlanego w panelu tekstowym oznaczenia na „SVEN” (*S*-treptomyces *VEN*-ezuelae). Oznacza to, że tryb pracy aplikacji został zmieniony na gatunek *S. venezuelae*, zarazem wyświetlany genom na głównym panelu zostanie zmieniony na genom tego gatunku. Następną opcją do wyboru jest „Show/hide visualizations”, w której możemy wybrać, które wykresy są wyświetlane. Domyślnie wyświetlany jest genom i wykres z danych RNA-seq i mikromacierzy. Trzecią opcją jest wykres z danych ChIP-seq, który należy włączyć, aby wyświetlany był na panelu głównym. Aby to zrobić należy kliknąć na pole tej opcji i wybrać opcję ChIPplot. Aby wyłączyć wyświetlanie danego wykresu, należy kliknąć na kafelek mu odpowiadający i nacisnąć przycisk „backspace” na klawiaturze. „Choose gene from list” pozwala na wybranie konkretnego interesującego nas genu. Gdy chcemy przybliżyć wykres na konkretny gen, należy kliknąć na panel, na którym widnieje „all” i wybrać interesujący nas gen. Następnie, aby zatwierdzić wybór należy nacisnąć przycisk „Apply Changes”. Wykres automatycznie zostanie ograniczony do wyświetlania wartości granicznych odpowiadających danemu genomu. Lista wyświetlanych genów zależy od wybranego gatunku.



Rysunek 4.3 Panel główny strony „RNA-seq & ChIP-seq visualization” po naciśnięciu przycisku „Apply Changes” znajdującego się w zakładce „Navigation” panelu bocznego. Wyświetlany jest wykres genomu wybranego gatunku.

Poniżej ustawień ogólnych, znajdziemy panel „RNA-seq Plot Data” i „ChIP-seq Plot Data”. Są to dwa panele odpowiadające za wybieranie danych z bazy dostępnej w aplikacji, które będą wyświetlone na wykresie. Dla wykresu RNA możemy wybrać do trzech zestawów danych z publikacji, natomiast dla ChIP-seq możliwe wybranie jest tylko jednego. Aby wybrać dane należy nacisnąć na „no data selected”. Spowoduje to rozwinięcie listy dostępnych danych dla danego gatunku, opisanych w [Tab. 3.5]. Po wybraniu jednej z opcji z listy, pojawi się pod spodem dodatkowe okno „Choose comparison for analysis”. Dla każdego zestawu danych w tym panelu możemy wybrać serie danych, które odpowiadają konkretnemu porównaniu, między innymi, rodzajowi podłoża, godzinie czy porównywanym szczepom. Wybrane przez nas dane, zostaną automatycznie wyświetlone na panelu głównym w postaci wykresu.

Następną zakładką opcji dostępną dla użytkownika jest „Plot settings” [Rys. 4.2 (C)]. Znajdują się w niej opcje modyfikacji wykresu na podstawie konkretnych wartości danych zmiennych. Dostępne są dwa filtry do użycia, logFC i FDR. Filtr logFC składa się z dwóch okien z wartościami numerycznymi, ustawionymi domyślnie na 0. W oknie „higher value” możemy ustawić dowolną liczbę większą od 0, natomiast w „lower value”, mniejszą od zera. Po zmianie którejkolwiek z wartości, wyświetlane wyniki zostaną odfiltrowane tak, aby geny które zobaczymy na wykresie znajdowały się poza zakresem logFC ustawionym przez

użytkownika. Drugim filtrem znajdującym się w zakładce „Plot settings” jest filtr FDR. FDR wyświetlanych wyników po włączeniu filtra zostanie przefiltrowane tak aby wyświetlać tylko wyniki mniejsze bądź równe 0.05. Taka wartość została wybrana ze względu na bardzo częste pojawianie się takiej granicy FDR jako punkt odcięcia danych przy wysokoprzepustowych badaniach NGS.

Zakładka „Upload data” służy użytkownikowi do załadowania własnych danych do programu [Rys. 4.2 (D)]. Przede wszystkim dane, które chcemy, żeby zostały wyświetlone należy przygotować do odpowiedniego formatu zanim zostaną dodane do aplikacji. Szablon formatu dla danych RNA-seq i mikromacierzy, jak i ChIP-seq znajduje się w wersji do pobrania na stronie „Help&Info” programu jak i w formie tabel w sekcji materiały i metody [Tab. 3.2, Tab. 3.3, Tab. 3.4]. Dane w formacie domyślnym dla programu pozwalają użytkownikowi na wpisanie jedynie nazw genów badanych, logFC i FDR, oraz opisanie ich wybraną przez siebie zmienną. W zależności od tego jaki gatunek wybrany jest w zakładce „Selection” taki genom zostanie przyrównany do danych użytkownika w celu dodania początków i końców genów na genomie. Dane załadowane do programu pozostają w nim tylko do zakończenia działania danej sesji, nie zostają one zapisane w nim na stałe. W „Upload data” widnieją dwie opcje dodania danych i przełącznik „Is RNA-seq file in app format?”, w zależności jaki typ danych chcemy załadować. Dane ChIP-seq mają tylko jedną opcję ładowania w formacie takim jak w Tabeli 3.3. Dane do RNA-seq mają dwie opcje ładowania. Domyślnie ustawiona jest ta w formacie danych w programie, takim jak opisany w Tabeli 3.2. Aby załadować dane RNA-seq, które nie są w formacie „tidy” należy przełączyć „Is RNA-seq file in app format?” na opcję „No”. Wtedy załadowane dane zostaną automatycznie przekształcone w format „tidy”. Dane dodajemy do programu poprzez naciśnięcie przycisku „Browse...” i wybranie pliku z naszego komputera, lub przez przeciągnięcie pliku na odpowiednie okno ładowania danych. Po załadowaniu wyświetli się okno, w którym użytkownik musi wpisać nazwę dla swojego pliku. Wpisana nazwa, wyświetlana będzie w opcjach wyboru danych w zakładce „Selection” i na wykresie wygenerowanym z danych użytkownika.

Opcją dostępną dla użytkownika w zakładce „Plot Download” jest pobranie wyświetlonych na panelu głównym wykresów w postaci pliku png [Rys. 4.2 (E)]. Wykres pobierany jest poprzez naciśnięcie przycisku „Download png plot”. Użytkownik ma też możliwość dostosować wysokość, szerokość i rozdzielczość wygenerowanego obrazu, do swoich potrzeb.

4.1.3 Wyświetlanie danych „RNA-seq & ChIP-seq visualization”

Gdy wyświetlony jest wykres genomu wybranej przez nas bakterii, możemy rozpocząć wybieranie interesujących nas danych. Działanie aplikacji w tym przypadku zostanie przedstawione na podstawie danych z publikacji badającej między innymi geny *satKR* dla RNA-seq [69]. Wybrane zostały interesujące nas publikacje i dostępne dla nich porównania [Rys. 4.4 (A)]. W ten sposób na panelu główny zaczynają wyświetlać się właściwe wykresy wybranych danych oznaczonych nazwami takimi jak wybrane w zakładce „Selection” [Rys. 4.5 (A)]. Następnie, gdy wyświetlone są już interesujące nas dane, w zakładce „Plot settings” można zmienić niektóre z wyświetlanych wartości:

- LogFC, za którego zmianę odpowiada „LogFC Filtering” [Rys. 4.4 (B)]. Wartości logFC zmienione zostały z 0 dla górnej granicy i 0 dla dolnej na 0,5 dla obu granic. Na panelu głównym, dane są aktualizowane w zależności o ustawione granice i wyświetlane już odfiltrowane [Rys. 4.5 (B)].
- FDR, za który odpowiada przełącznik „Leave only significant genes” [Rys. 4.4 (B)]. Na panelu głównym wyświetlane są tylko wyniki dla FDR mniejsze bądź równe 0,05. Widoczny wykres prezentuje odfiltrowane wyniki fałszywie pozytywne [Rys. 4.5 (C)].

Jak zostało opisane wcześniej, w zakładce „Upload data” po załadowaniu własnych danych i nadaniu im nazwy, są one dostępne do wyboru w zakładce „Selection” [Rys. 4.4 (C)]. Po ich wybraniu, dla tego przykładu dane zostały nazwane „dane_test_2025”, zaobserwujemy je wyświetlone na wykresie panelu głównego, wraz z widniejącą wpisaną przez nas nazwą [Rys. 4.5 (D)]. Gdy ładujemy własne dane, program musi zaktualizować listę załadowanych danych. Z tego powodu, wcześniej ustawione opcje w zakładce „Selection” w podzakładkach wyborów danych RNA-seq i ChIP-seq, zostaną zresetowane do opcji „no data selected”. Dlatego zaleca się załadowanie własnych danych do aplikacji jako pierwszą czynność, przed dalszą analizą.

RNA-seq Plot Data **A**

SatKR_Gongerowska_2021

Choose contrasts for analysis

M145vsMGM10

M145vsMGM12

M145vsMGM11

M145vsMGM14

data_test_2025

Choose contrasts for analysis

variable_1

variable_2

no data selected

ChIP-seq Plot Data

SMChupS_Szafran_2021

Choose contrasts for analysis

SMC_14

LogFC Filtering **B**

-higher value

0,5

-lower value

-0,5

☐

Leave only significant genes (FDR <= 0.05)

OFF

RNA-seq **C**

Browse...

dane_user_new.txt

Upload complete

Add name for your file

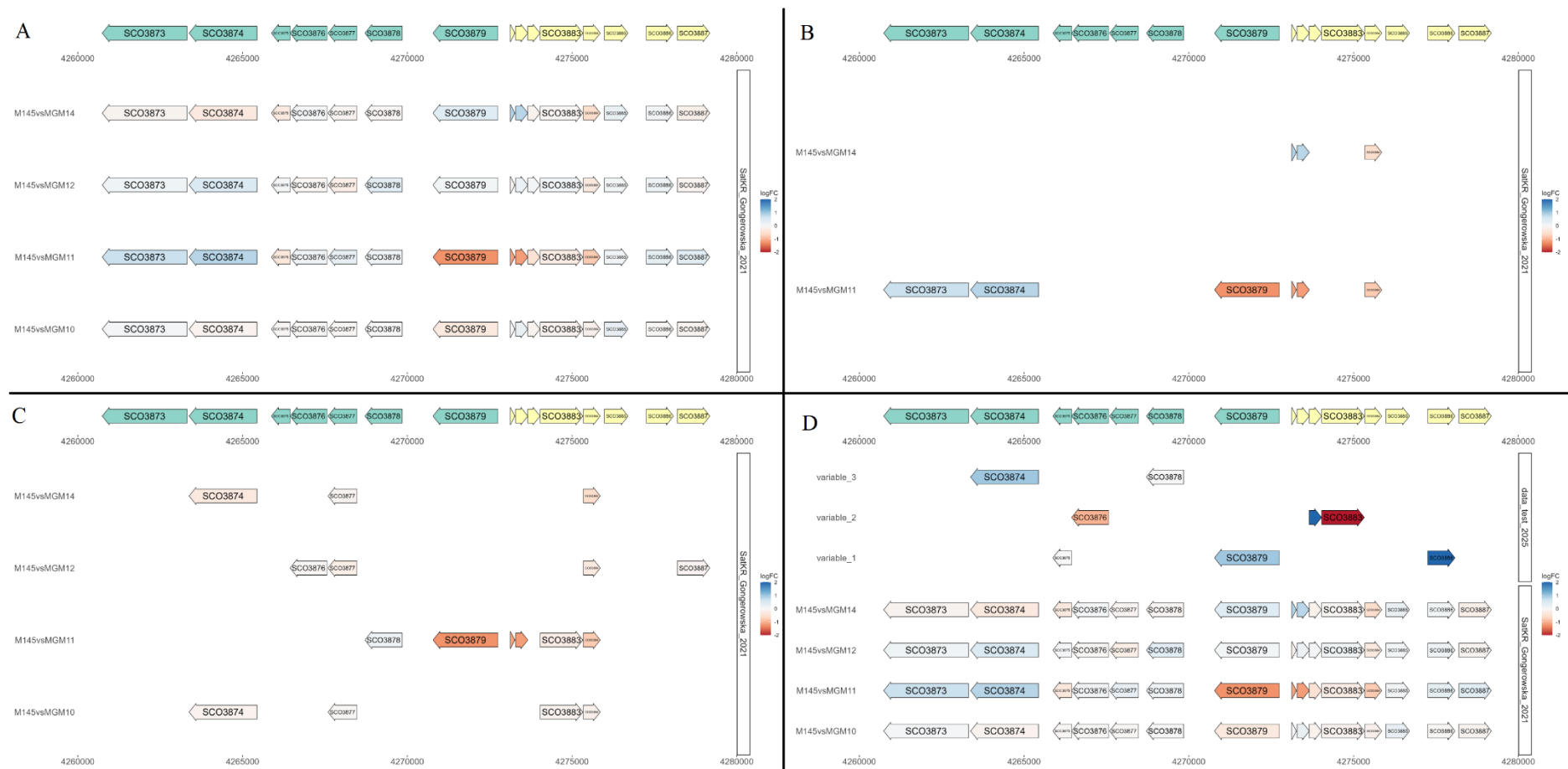
data_test_2025

CHIP-seq

Browse...

No file selected

Rysunek 4.4 Zakładki panelu bocznego „RNA-seq & ChIP-seq visualization” z wybranymi przykładowymi danymi. Zakładka „Selection” z wybranymi danymi (A), zakładka „Plot settings” ze zmienioną wartością granic logFC (B), dane użytkownika dodane w zakładce „Upload Data” (C).



Rysunek 4.5 Panel główny „RNA-seq & ChIP-seq visualization” po wybraniu przykładowych danych (A), zmianie wartości wyświetlanego logFC na wykresie (B), włączeniu filtra FDR (C), załadowaniu i wyświetleniu własnych danych (D).

Dla użytkownika w aplikacji, dostępna jest również lista wyświetlonych na wykresie danych w tabelach [Rys. 4.6]. Tabele danych wykresów RNA-seq i ChIP-seq znajdują się pod wykresami panelu głównego w dwóch zakładkach „RNAseq on plot”, „RNAseq no filter”, „ChIPseq on plot” i „ChIPseq no filter”. Tabela „RNAseq on plot” zawiera tylko dane dotyczące tych genów, które w danej chwili widoczne są na wykresie strzałkowym RNA-seq, natomiast „RNAseq no filter” pokazuje wszystkie dane które dostępne są dla wybranych zbiorów danych i porównań. „ChIPseq on plot” wyświetla tylko dane widniejące na wykresach ChIP-seq, a „ChIPseq no filter”, odpowiednio do „RNAseq no filter”. Wyświetlane tabele można przeszukiwać poprzez wpisywanie interesujących nas haseł do „Search:” znajdującego się nad tabelami.

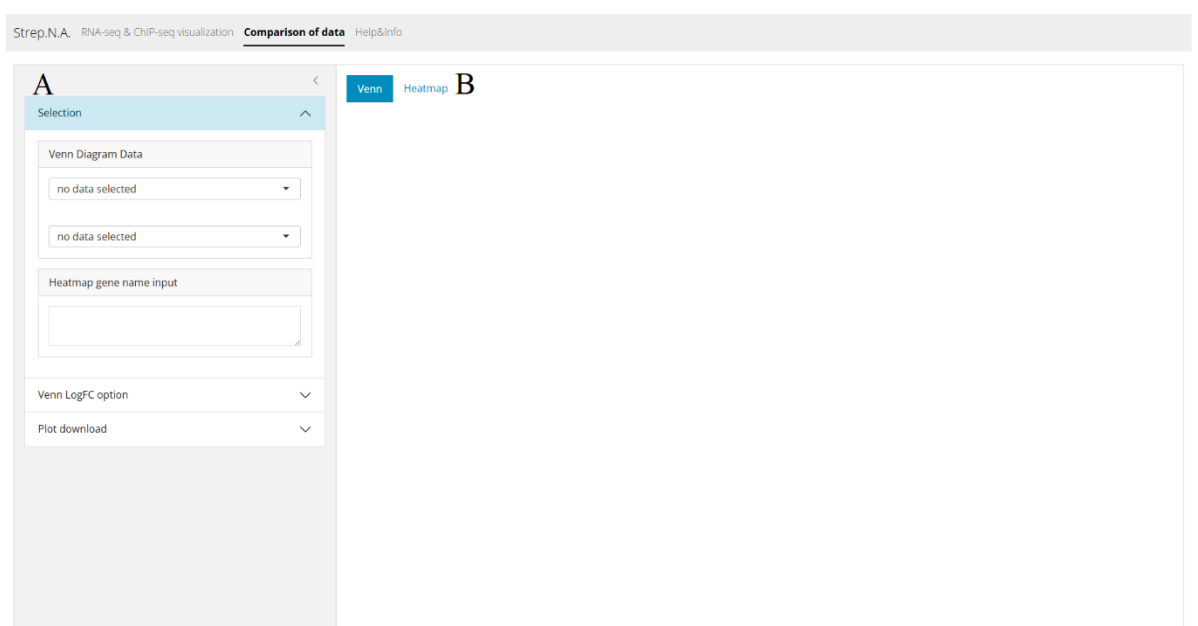
RNAseq on plot RNAseq no filter ChIPseq on plot ChIPseq no filter									
Show 10 entries		Search:							
	gene	FDR	logFC	add_variable	start	end	width	strand	data_name
1	SCO3885	1	-0.29	MGM10vsM145	4275980	4276699	720	+	SatKR_Gongerowska_2021
2	SCO3873	0.061932977	-0.07	MGM10vsM145	4260741	4263314	2574	-	SatKR_Gongerowska_2021
3	SCO3878	0.061750963	0	MGM10vsM145	4268713	4269843	1131	-	SatKR_Gongerowska_2021
4	SCO3881	0.419912987	-0.27	MGM10vsM145	4273286	4273657	372	+	SatKR_Gongerowska_2021
5	SCO3880	0.30167841	-0.01	MGM10vsM145	4273127	4273264	138	+	SatKR_Gongerowska_2021
6	SCO3873	0.512937001	-0.71	MGM11vsM145	4260741	4263314	2574	-	SatKR_Gongerowska_2021
7	SCO3874	0.796828989	-0.84	MGM11vsM145	4263378	4265438	2061	-	SatKR_Gongerowska_2021
8	SCO3876	0.066505914	-0.13	MGM11vsM145	4266442	4267563	1122	-	SatKR_Gongerowska_2021
9	SCO3877	0.089646732	-0.28	MGM11vsM145	4267594	4268469	876	-	SatKR_Gongerowska_2021
10	SCO3878	0.04372753	-0.13	MGM11vsM145	4268713	4269843	1131	-	SatKR_Gongerowska_2021
Showing 1 to 10 of 61 entries					Previous 1 2 3 4 5 6 7 Next				

Rysunek 4.6 Tabele przedstawiają dane wybrane w zakładce „Selection”. Tabele oznaczone jako „...on plot” zawierają dane aktualnie wyświetlane na wykresie, natomiast tabele „...no filter” prezentują pełny zestaw danych, bez ograniczeń wynikających z filtrów osi wykresu.

4.1.4 Panel boczny i funkcje „Comparison of data”

Po naciśnięciu „Comparison of data” program zostaje przełączony na stronę dotyczącą porównań między danymi. Strona ta, podobnie jak strona „RNA-seq & ChIP-seq visualization” składa się z panelu bocznego, gdzie możemy dokonywać wyborów danych i opcji [Rys. 4.7 (A)], jak i z panelu głównego, gdzie będą wyświetlane wykresy [Rys. 4.7 (B)]. Panel boczny tej strony składa się z czterech zakładek, natomiast panel główny z trzech.

Zakładka „Selection” jest zakładką umożliwiającą wybór danych dla strony „Comparison of data” [Rys. 4.8 (A)]. W tej zakładce znajdziemy dwie podzakładki. Pierwsza z nich „Venn Diagram Data”, służy do wyboru danych, które będą wyświetlane na panelu głównym dla wykresów Venn’a i Heatmapy. Po naciśnięciu „no data selected” rozwinię się lista zawierająca wszystkie dane dostępne w aplikacji Podobnie jak przy wyborze danych na pierwszej stronie programu, po wybraniu konkretnej publikacji, należy wybrać, które porównania chcemy zobaczyć na wykresie.



Rysunek 4.7 Strona „Comparison of data”. Widoczny jest boczny panel przycisków (A) i panel główny zawierający 3 podstrony, na którym wyświetlać się będą wykresy (B).

W dostępnej wersji aplikacji diagram Venna funkcjonuje poprawnie dla maksymalnie 4 zbiorów danych, gdzie jako zbiór danych określa się jedno porównanie. Zatem do analizy wybranych może zostać np.: cztery porównania z jednej publikacji bądź po dwa porównania z dwóch publikacji. Gdy wybrane zostanie powyżej czterech porównań diagram Venna, przekształcany jest w wykres typu „upset”. Przedstawia on więcej zestawień między porównaniami w formie wykresu słupkowego obrazującego liczbę wyników wspólnych między różnymi zbiorami. Pod wykresem słupkowym pokazane jest jakie zestawienie odpowiada danemu słupkowi poprzez zaznaczenie ich kropką. Druga pod-zakładką jest „Heatmap gene name input”. W oknie tekstowym tej podzakładki wpisać można nazwy interesujących nas genów w formacie nazw genów wpisanych po przecinku na przykład: SCO6276, SCO6277, SCO6278, SCO6279, SCO6280. Wpisane geny zostaną odfiltrowane

z wybranych w pierwszej podzakładce danych i wyświetlone w formie Heatmapy w zakładce „Heatmap” panelu głównego.

Zakładka „Venn LogFC option” służy do zmiany wartości logFC danych użytych do przygotowania diagramu Venna [Rys. 4.8 (B)]. Należy podkreślić, że nie wpływa on na wartości logFC wyświetlające się na Heatmapie. Podobnie jak na pierwszej stronie programu, wybrane przez nas dane są filtrowane na podstawie ustawionej górnej i dolnej granicy logFC. Domyślnymi wartościami w górnej i dolnej granicy ustawionymi w tej zakładce jest 1,5 i -1,5. Strona „Comparison of data” nie zawiera przełącznika dla FDR. Wszystkie dane wyświetlane na tej stronie są domyślnie filtrowane dla FDR mniejszego bądź równego 0,05.

Ostatnia zakładka „Plot download” podobnie jak na stronie pierwszej programu służy do pobierania wyświetlanych wykresów w formacie „png” [Rys. 4.8 (C)]. Oddzielnie znajduje się na niej opcja pobrania diagramu Venn’a, pod „Plot Download Venn” i pod nią opcja pobrania Heatmapy „Plot Download Heat”.

The image shows a sidebar with three panels. The top panel, 'Selection A', contains two dropdown menus for 'Venn Diagram Data' (both set to 'no data selected') and a text input for 'Heatmap gene name input'. The middle panel, 'Venn/Heat LogFC B', has input fields for 'higher logFC' (1,5) and 'lower logFC' (-1,5). The bottom panel, 'Plot download C', contains two sections: 'Plot Download Venn' and 'Plot Download Heat'. Each section has a 'Download png plot' button and input fields for 'Plot width [cm]' (20), 'Plot height [cm]' (14), and 'Resolution' (200).

Rysunek 4.8 Panel boczny strony „Comparison of data” z rozwiniętymi zakładkami „Selection” (A), „Venn/Heat LogFC” (B), „Plot download” (C). (w orientacji poziomej).

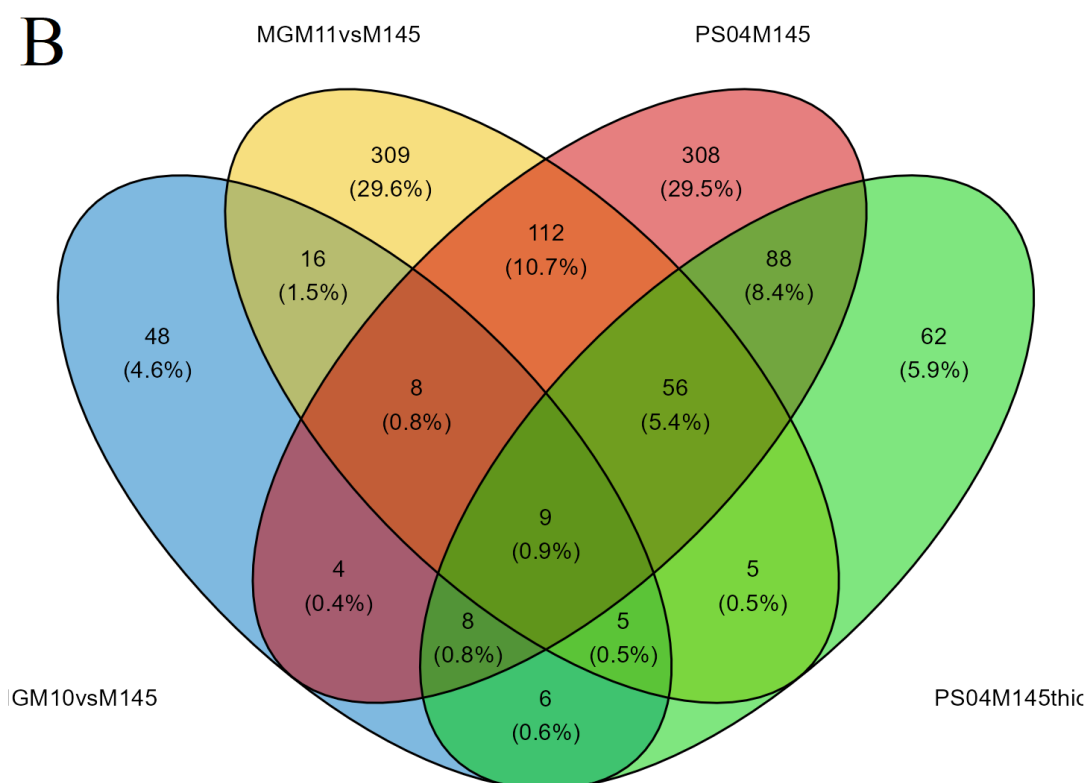
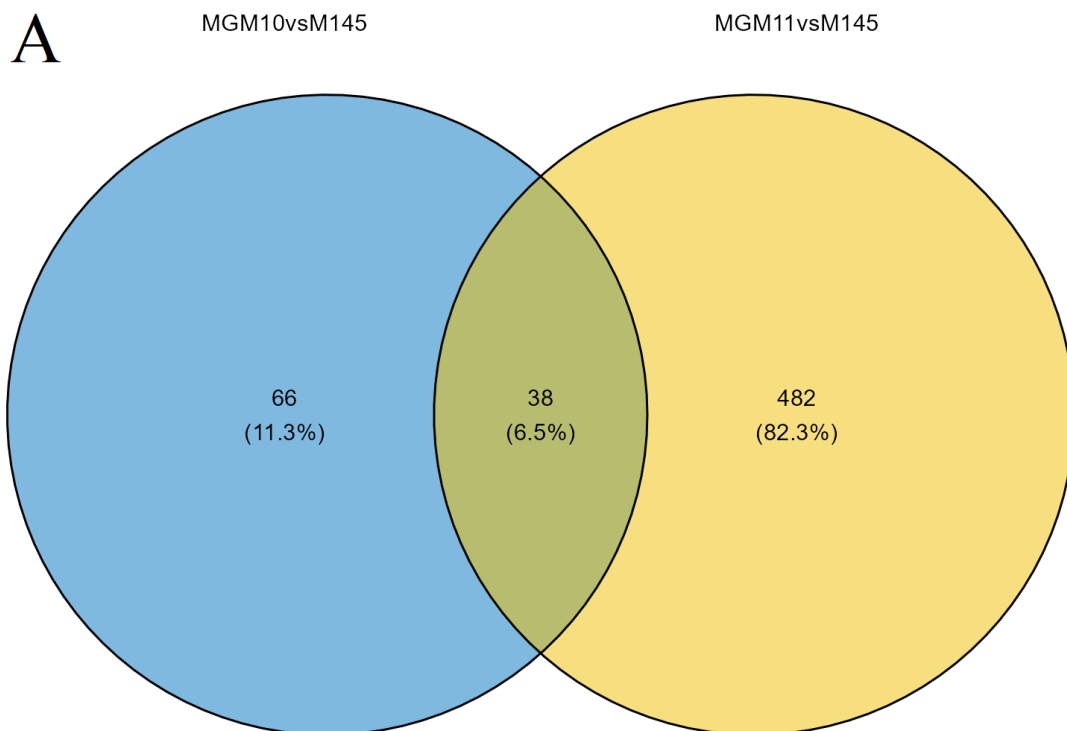
4.1.5 Wyświetlanie danych „Comparison of data”

Jak zostało wspomniane wcześniej, możemy na wykresie Venn’a porównywać do maksymalnie czterech zbiorów. Aby wyświetlić wykres na panelu głównym należy wybrać kilka porównań dla jednego eksperymentu [Rys. 4.9 (A)], bądź rozdzielić cztery dostępne na dwa eksperymenty [Rys. 4.9 (B)].

The image shows a web interface with two panels, A and B, under the 'Selection' tab. Both panels have a 'Venn Diagram Data' section. Panel A shows a dropdown menu with 'SatKR_Gongerowska_2021' selected, and a 'Choose comparisons for venn' section with two buttons: 'MGM10vsM145' and 'MGM11vsM145'. Below this is a dropdown menu with 'no data selected'. Panel B shows the same dropdown menu for 'SatKR_Gongerowska_2021' with the same two comparison buttons. Below this is a dropdown menu with 'TopA_Szafran_2019' selected, and another 'Choose comparisons for venn' section with two buttons: 'PS04M145' and 'PS04M145thio'. Both panels have a 'Heatmap gene name input' section at the bottom.

Rysunek 4.9 Zakładka „Selection” strony „Comparison of data” z wybranymi przykładowymi danymi. Dane wybrane w celu zestawienia 2 porównań dla jednego eksperymentu (A); dane wybrane w celu zestawienia po dwa porównania dla 2 zestawów danych (B).

Po wybraniu interesujących nas danych na panelu głównym wyświetli się diagram Venna. Przykładowo, dla dwóch porównań pochodzących z jednej publikacji, zobaczymy porównanie z jedną częścią wspólną dla dwóch zbiorów [Rys. 4.10 (A)], natomiast dla dwóch publikacji, gdzie wybrane dla każdej z nich zostało po dwa porównania, wyświetlony wykres będzie składał się z 4 zbiorów i zawierał części wspólne między poszczególnymi zbiorami, jak i dla ich wszystkich [Rys. 4.10 (B)].

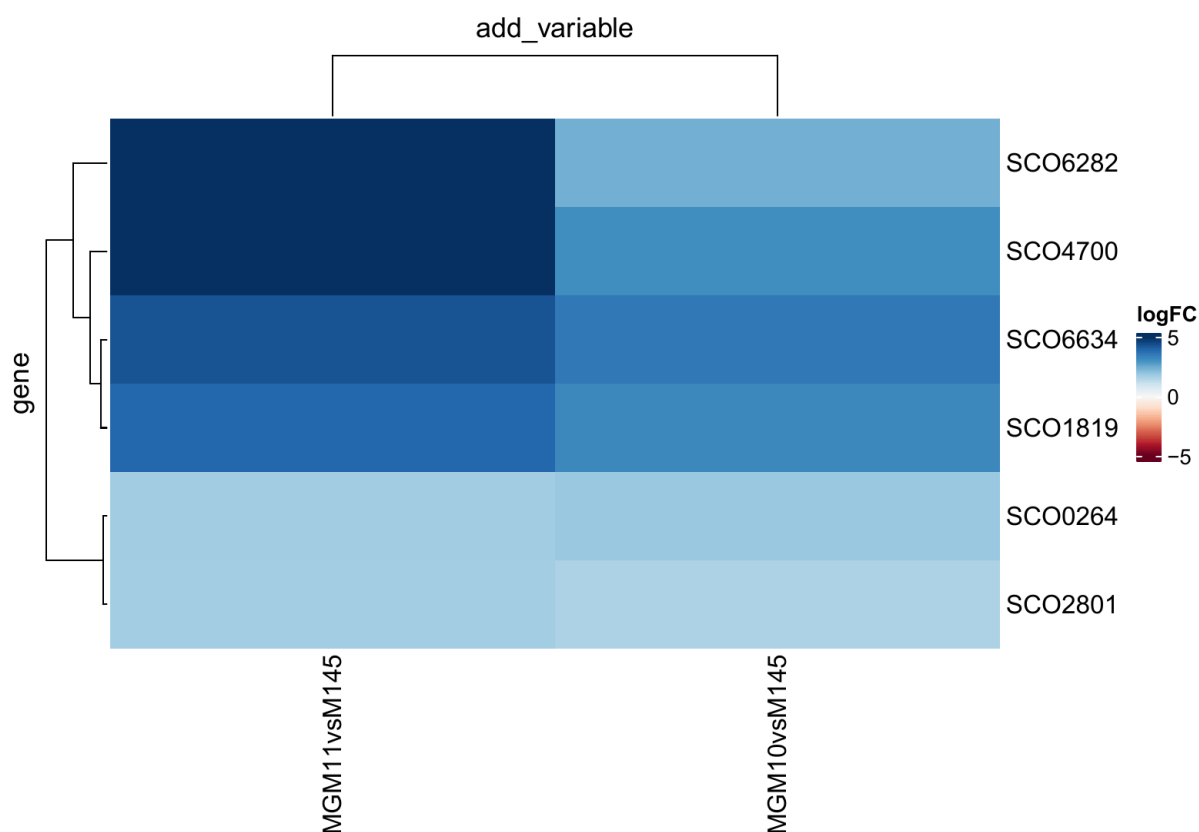


Rysunek 4.10 Diagram Venna wyświetlany na podstawie wybranych danych w celu zestawienia 2 porównań dla jednego eksperymentu (A); dane wybrane w celu zestawienia po dwie porównań dla 2 zestawów danych (B).

Poniżej diagramu Venna znajdują się geny wspólne i specyficzne tylko dla niektórych podzbiorów [Rys. 4.11]. Geny wyświetlające się pod diagramem Venna można skopiować i wkleić do panelu tekstowego „Heatmap gene name input”. Po wklejeniu genów do panelu tekstowego, poprzez naciśnięcie zakładki „Heatmap” panelu głównego przejdziemy do Heatmapy powstałej z wybranych przez nas danych [Rys. 4.12]. Dokładniej Heatmapa tworzy się na podstawie wybranych przez nas eksperymentów, dane odfiltrowywane są na podstawie wybranych porównań, a ostatecznie przez wybrane przez nas geny. Geny wklejane do panelu tekstowego „Heatmap gene name input”, nie muszą być zaczerpnięte z diagramu Venna, może to być dowolna lista genów wklejona w odpowiednim formacie. Heatmapa grupuje również podobne geny i porównania w formie drzewek, z lewej strony dla podobnych genów, a u góry dla podobnych porównań. Gdy wybranego na heatmapie genu brakuje w zbiorze jest on zaznaczany na kolor szary.

Show	10	entries	Search:	
Combination	Elements	A		
1	Common to: M145vsMGM10, M145vsMGM11	SCO0573, predicted RNA, SCO2864, SCO4667, SCO0572, SCO4686, SCO0571, SCO4214, SCO6205, SCO3426, SCO0666, SCO5957, SCO1293, SCO4669, SCO0985, SCO2865, SCO0575, SCO5226, SCO5035, SCO107, SCO1819, SCO6634, SCO2801, SCO4700, SCO0264, SCO6282, SCO3690, SCO46, SCO4449, SCO6299, SCO0772, SCO138, SCO5821, SCO121, SCO6800, SCO1364, SCO3089, SCO5254		
Showing 1 to 1 of 1 entries			Previous	1
Show	10	entries	Search:	
Combination	Elements	B		Count
1	Unique to: M145vsMGM10	SCO0773, SCO1902, SCO0861, SCO3281, SCO6481, SCO0436, SCO4674, SCO7440, SCO0986, SCO1570, SCO0863, SCO5105, SCO5959, SCO3427, SCO0574, SCO4682, SCO7747, SCO7252, SCO4681, SCO5582, SCO4945, SCO4680, SCO4678, SCO3897, SCO4215, SCO2717, SCO0915, SCO6514, SCO5410, SCO0832, SCO6804, SCO4564, SCO0831, SCO1256, SCO4006, SCO7186, SCO1541, SCO4166, SCO2258, SCO4332, SCO2927, SCO6091, SCO5473, SCO3110, SCO0597, SCO3111, SCO4586, SCO3390, SCO0973, SCO4585, SCO7646, SCO7536, SCO1366, SCO4317, SCO4005, SCO3958, SCO3959, SCO6509, SCO3957, SCO3956, SCO3090, SCO4903, SCO4584, SCO1550, SCO0834, SCO503		66
2	Unique to: M145vsMGM11	SCO0137, SCO0218, SCO0409, SCO0421, SCO0459, SCO0489, SCO0490, SCO0492, SCO0498, SCO0499, SCO0560, SCO0607, SCO0616, SCO0617, SCO0722, SCO0762, SCO0799, SCO0826, SCO0839, SCO0911, SCO1096, SCO1198, SCO1235, SCO1442, SCO1481, SCO1482, SCO1484, SCO1567, SCO1690, SCO1785, SCO1786, SCO1878, SCO2018, SCO2270, SCO2271, SCO2272, SCO2273, SCO2274, SCO2444, SCO2483, SCO2633, SCO2782, SCO2783, SCO2784, SCO2785, SCO2897, SCO2996, SCO3026, SCO3108, SCO3169, SCO3339, SCO3811, SCO3835, SCO3964, SCO3965, SCO3966, SCO4048, SCO4049, SCO4050, SCO4055, SCO4134, SCO4142, SCO4145, SCO4762, SCO4773, SCO4916, SCO5034, SCO5036, SCO5114, SCO5225, SCO5308, SCO5316, SCO5402, SCO5444, SCO5484, SCO5497, SCO5500, SCO5530, SCO5655, SCO5656, SCO5693, SCO5837, SCO5862, SCO5863, SCO5935, SCO6471, SCO6501, SCO6503, SCO6504, SCO6505, SCO6506, SCO6520, SCO6550, SCO6591, SCO6705, SCO6706, SCO6731, SCO6765, SCO7108, SCO7131, SCO7132, SCO7227, SCO7233, SCO7315, SCO7316, SCO7363, SCO7398, SCO7399, SCO7400, SCO7412, SCO7426, SCO7428, SCO7443, SCO7452, SCO7453, SCO7465, SCO7469, SCO7470, SCO7471, SCO7472, SCO7473, SCO7474, SCO7475, SCO7657, SCO7658, SCO0063, SCO0075, SCO0091, SCO0102, SCO0175, SCO0289, SCO0358, SCO0359, SCO0368, SCO0408, SCO0433, SCO0462, SCO0463, SCO0484, SCO0485, SCO0512, SCO0523, SCO0550, SCO0558, SCO0566, SCO0630, SCO0631, SCO0646, SCO0648, SCO0714, SCO0725, SCO0729, SCO0758, SCO0781, SCO0800, SCO0883, SCO0901, SCO0909, SCO0916, SCO1000, SCO1003, SCO1069, SCO1077, SCO1083, SCO1104, SCO1191, SCO1307, SCO1308, SCO1313, SCO1321, SCO1322, SCO1339, SCO1397, SCO1406, SCO1425, SCO1448, SCO1454, SCO1455, SCO1471, SCO1497, SCO1540, SCO1584, SCO1588, SCO1608, SCO1624, SCO1625, SCO1729, SCO1736, SCO1779, SCO1790, SCO1791, SCO1795, SCO1904, SCO1905, SCO1926, SCO1927, SCO1979, SCO1986, SCO1987, SCO1988, SCO2006, SCO2027, SCO2223, SCO2224, SCO2248, SCO2264, SCO2351, SCO2379, SCO2459, SCO2460, SCO2466, SCO2485, SCO2496, SCO2498, SCO2525, SCO2537, SCO2551, SCO2632, SCO2652, SCO2654, SCO2679, SCO2697, SCO2727, SCO2763, SCO2768, SCO2815, SCO2876, SCO2894, SCO2895, SCO2896, SCO2898, SCO2964, SCO2986, SCO3001, SCO3035, SCO3207, SCO3258, SCO3264, SCO3268, SCO3269, SCO3292, SCO3299, SCO3329, SCO3331, SCO3332, SCO3360, SCO3389, SCO3423, SCO3424, SCO3447, SCO3492, SCO3508, SCO3509, SCO3595, SCO3604, SCO3608, SCO3616, SCO3631, SCO3640, SCO3657, SCO3708, SCO3709, SCO3766, SCO3773, SCO3782, SCO3867, SCO3972, SCO3975, SCO3982, SCO3983, SCO3984, SCO3985, SCO3986, SCO3987, SCO3988, SCO4011, SCO4056, SCO4072, SCO4085, SCO4096, SCO4107, SCO4108, SCO4117, SCO4121, SCO4125, SCO4183, SCO4222, SCO4264, SCO4278, SCO4283, SCO4316, SCO4336, SCO4338, SCO4342, SCO4370, SCO4375, SCO4384, SCO4385, SCO4412, SCO4424		482

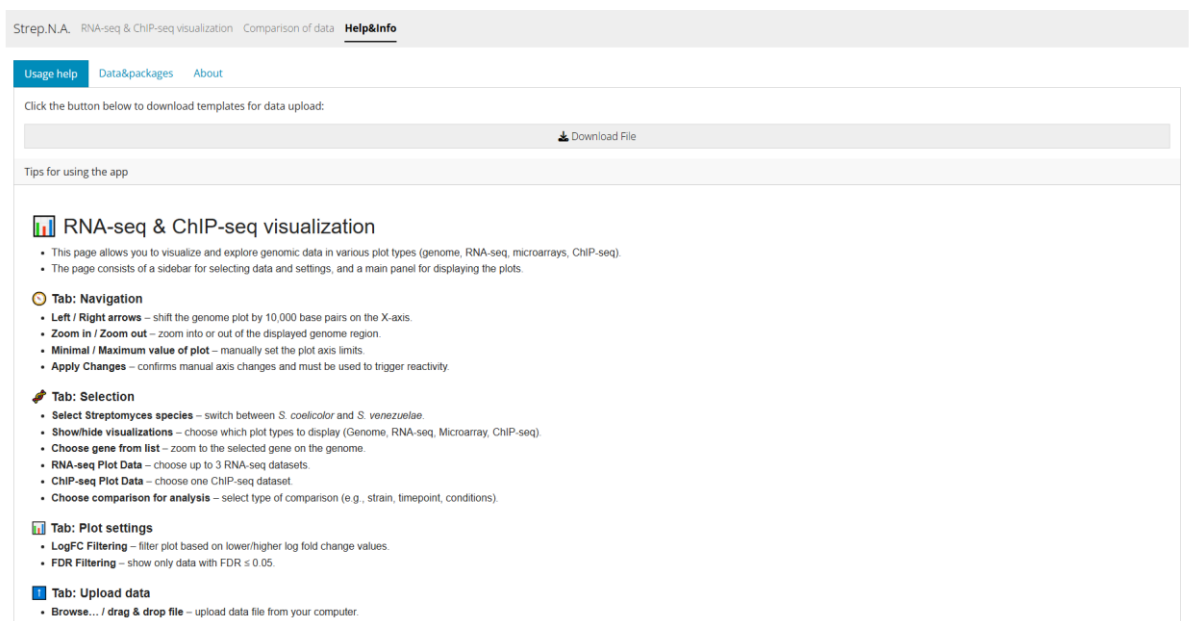
Rysunek 4.11 Geny wspólne (A) dla wyświetlonych danych na diagramie Venna, przedstawione w formie tabeli pod diagramem. Geny obecne tylko w danych zbiorach (B).



Rysunek 4.12 Heatmapa powstała z pierwszych sześciu genów wspólnych dla wybranych danych, na podstawie diagramu Venna.

4.1.6 Strona „Help&Info”

Ostatnią stroną programu Strep.N.A, jest „Help&Info” [Rys. 4.13]. Znajdują się na niej wszystkie najważniejsze informacje dotyczące użytkowania programu jak i wykorzystywanych w nim danych czy pakietów. Strona „Help&Info” składa się z 3 zakładek „Usage help”, „Data&Packages” i „About”. Na pierwszej zakładce znajdziemy wszelkie informacje dotyczące jak korzystać z aplikacji. Dostępny dla użytkownika jest szablon przedstawiający, jak powinny wyglądać dane dodawane do aplikacji w formie pliku do pobrania. Druga zakładka składa się z dwóch tabel. Pierwsza zawiera informacje dotyczące pakietów R w wykorzystanych do stworzenia aplikacji, a druga informacje na temat publikacji, z których pochodzą dane znajdujące się w aplikacji. W trzeciej zakładce strony „Help&Info” znajdują się informacje kontaktowe do autorów aplikacji.



Rysunek 4.13 Pierwsza zakładka strony „Help&Info” programu.

4.2 Zastosowanie aplikacji

Na podstawie danych zamieszczonych w aplikacji, jak i opcji wprowadzania własnych danych dokonać można analiz różnych eksperymentów w celu sprawdzenia zależności jakie zachodzą między ekspresją genów w różnych szczepach. W aplikacji znajdują się dane z eksperymentów służących do pomiaru ekspresji genów takie jak RNA-seq i mikromacierze, jak i dane z eksperymentów służących do oznaczania miejsc przyłączania się białek typu ChIP-seq. Poprzez analizę różnic między eksperymentami jesteśmy w stanie wykonać takie porównania jak zestawienie dwóch pomiarów RNA-seq, czy wyników RNA-seq z ChIP-seq. Na podstawie takich analiz można wyciągnąć wnioski dotyczące zależności między różnymi genami, między innymi ich funkcje regulatorowe.

4.2.1 RNA-seq i ChIP-seq

RNA-seq pozwala na pomiar ekspresji genów, natomiast ChIP-seq umożliwia odczytanie, gdzie w genomie wiążą się interesujące badacza białka. Poprzez zestawienie ze sobą wyników z tych dwóch technik można określić wiele zależności takich jak wpływ obecności danego białka na ekspresję określonych genów lub za jak szeroką regulację genomu odpowiada dany czynnik. Podczas zestawienia ze sobą tych dwóch technik możemy również stwierdzić, czy białko kontroluje ekspresję danych genów w sposób bezpośredni, poprzez wiązanie się do konkretnego miejsca na genomie, czy pośredni, na przykład poprzez regulację innych białek, które będą się wiązały do genomu.

W przedstawionym poniżej przypadku, badane były dwa szczepy *S. venezuelae*. Jeden z nich był typem dzikim bakterii, a drugi posiadał delecję genu *adpA*. Na szczepach zostały przeprowadzone badania RNA-seq w 13 i 20 godzinie wzrostu, jak i ChIP-seq dla miejsc wiązania AdpA również w tych godzinach wzrostu bakterii (dane nieopublikowane, otrzymane dzięki uprzejmości dr. inż. Marcina Wolańskiego z Zakładu Mikrobiologii Molekularnej, wydziału biotechnologii, Uniwersytetu Wrocławskiego). Dane prezentowane poniżej nie znajdują się w bazie danych aplikacji i zostały ręcznie wczytane do programu dzięki funkcji „Upload data”.

AdpA jest ważnym regulatorem transkrypcyjnym u *Streptomyces*, ponieważ odpowiedzialny jest on jednocześnie za kontrolę różnicowania morfologicznego, jak i biosyntezę różnych związków chemicznych między innymi chloramfenikolu, czyli antybiotyku, którego działanie polega na hamowaniu syntezy białek bakteryjnych poprzez wiązanie z podjednostką 50S rybosomu [87], [88]. Badanie zależności między obecnością białka regulatorowego i klastrów genów produkujących antybiotyki, pozwala na zrozumienie regulacji procesów biosyntezy takich związków, a zatem na przykład na zoptymalizowanie wydajności biosyntezy co może pomóc w produkcji funkcjonalnych antybiotyków. Dlatego w celu zobrazowania działania aplikacji został wybrany klaster genów biosyntezy chloramfenikolu [Rys. 4.14]. W skład tego klastra wchodzi około 40 genów od *vnz_4355* do *vnz_4560*. Na podstawie wykresu ChIP-seq można zauważyć, że AdpA, w rejonie klastra chloramfenikolu, przyłącza się na jego początku. Sugeruje to, że faktycznie to białko regulatorowe ma wpływ na biosyntezę chloramfenikolu. Na podstawie wykresu RNA-seq możemy zauważyć, że różnica w ekspresji między dwoma szczepami była nieznaczna w 13 godzinie hodowli, natomiast w 20 godzinie, wartości logFC genów należących do klastra biosyntezy chloramfenikolu mają zdecydowanie zaniżony poziom ekspresji w stosunku do typu dzikiego. Produkcja AdpA nasila się, gdy bakterie kończą fazę wzrostu wegetatywnego i rozpoczynają tworzenie strzępek powietrznych, a fragmentacja strzępek następuje około 20 godziny hodowli [87]. Można zatem stwierdzić, że w porównaniu RNA-seq w 20 godzinie między szczepami AdpA powinno odgrywać największą rolę. Zatem niedobór tego białka powoduje zaniżoną ekspresję chloramfenikolu u szczepu delecyjnego co oznacza, że AdpA jest induktorem biosyntezy chloramfenikolu.

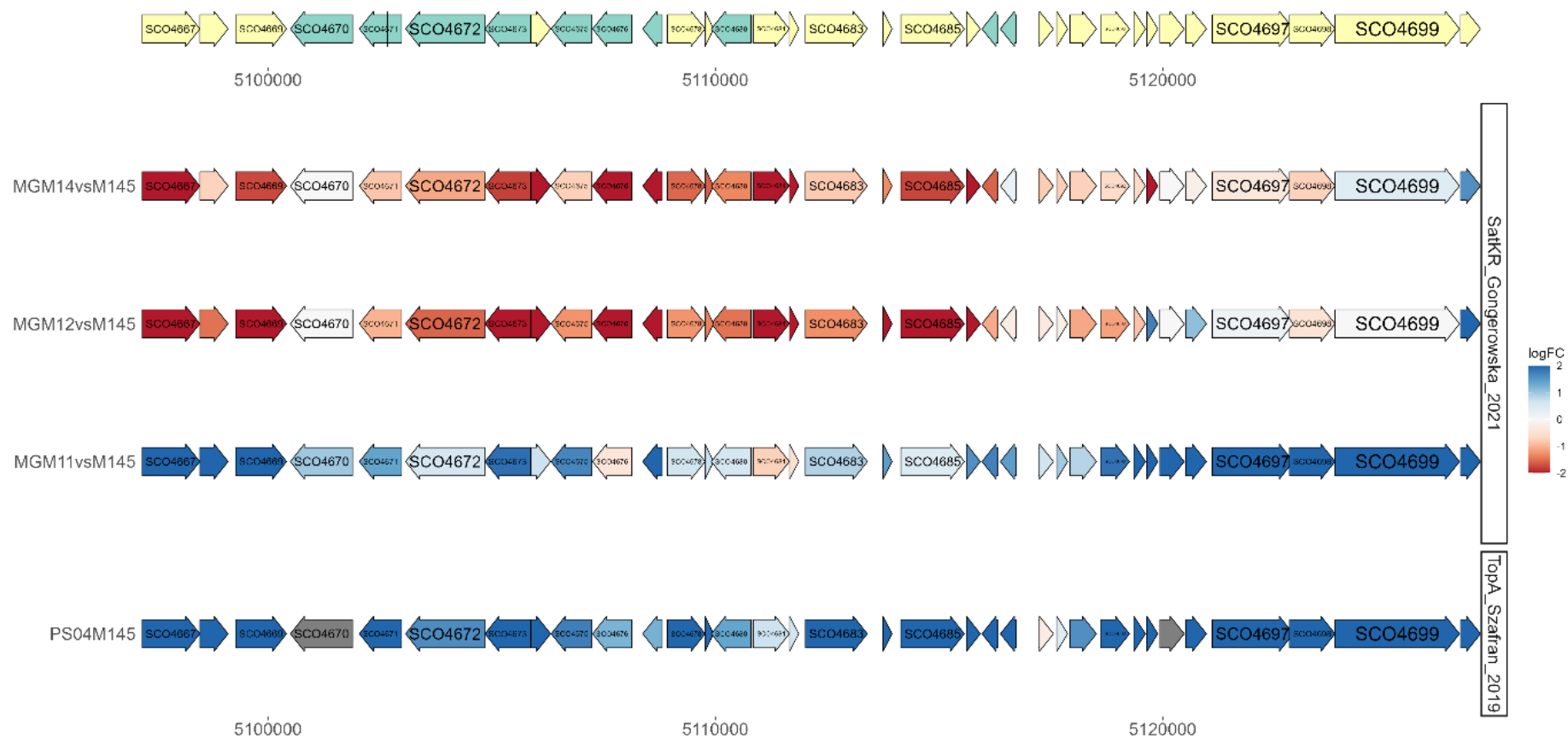


Rysunek 4.14 Wykres powstały przy użyciu aplikacji dla szczepu *S. venezuelae* z delecją genu *adpA*, przedstawiający klastery biosyntezy chloramfenikolu.

4.2.2 RNA-seq i RNA-seq

Innym zastosowaniem aplikacji jest użycie jej do porównania dwóch eksperymentów RNA-seq, które skupiały się na badaniu podobnych zależności czy regionów genomu. Dzięki zestawieniu wyników ekspresji dla dwóch eksperymentów, gdzie na przykład użyte były szczepy charakteryzujące się takimi samymi mutacjami, lecz wprowadzono inne warunki eksperymentu, można stwierdzić jak różne czynniki wpływają na dany klaster genów i dokładniej zrozumieć jego działanie.

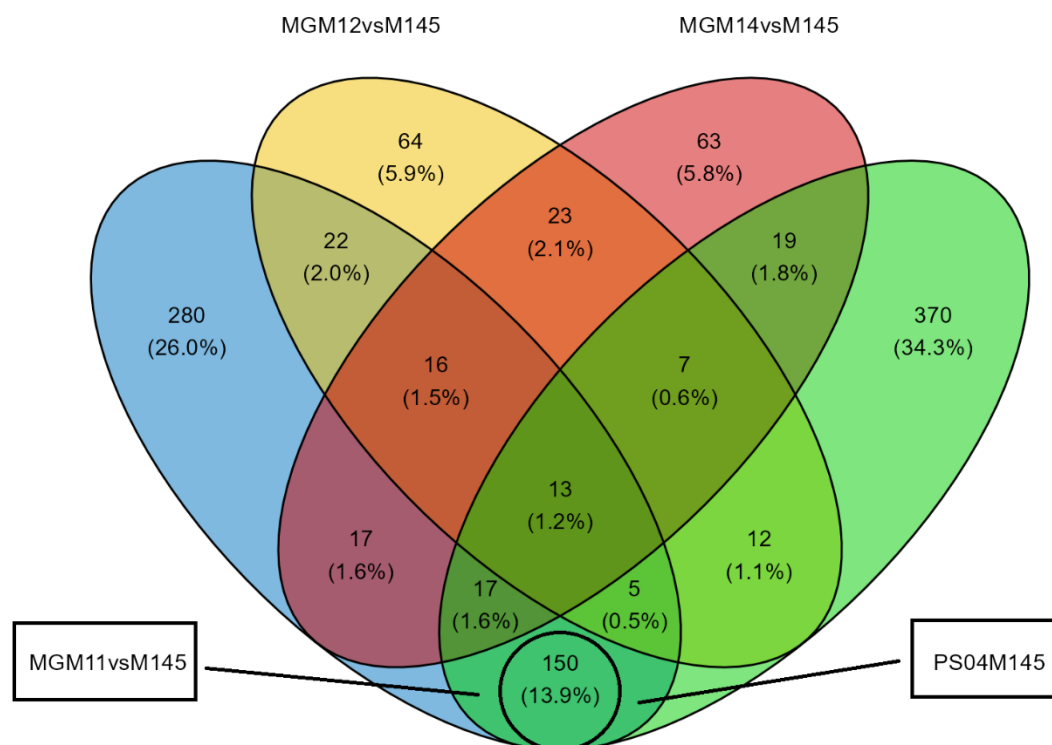
TopA to topoizomeraza, której główne działanie u *S. coelicolor* polega na usuwaniu superskrętów DNA w genomie bakterii, co jest szczególnie ważnym procesem podczas, między innymi, replikacji DNA w celu usuwania negatywnych superskrętów powstających podczas tego procesu, jak i transkrypcji genów [68]. Badania związane z obserwacją, jak genom bakterii reaguje na obecność TopA, prowadzą do zrozumienia aranżacji genomu i tego jak, bądź w jakich warunkach, zachodzi ekspresja niektórych genów. Dzięki zastosowaniu szczepów bakterii, w których obecność TopA jest zaniżona, udało się zidentyfikować region wyjątkowo czuły na warunki długotrwałego superskręcenia DNA, nazwany SHC (ang. supercoiling-hypersensitive region) (*sco4667* - *sco4700*) [68]. Na wykresie przygotowanym dzięki aplikacji zostały przedstawione wyniki związane z określeniem SHC jako istotnego klastra w *S. coelicolor* [Rys. 4.15]. Wyniki opisane jako „TopA_Szafran_2019”, przedstawiają porównanie szczepów *S. coelicolor*, M145, czyli typ dziki i PS04, czyli szczep z obniżonym poziomem TopA. Zauważyć można, że porównanie tych dwóch szczepów przedstawia zdecydowanie zwiększoną ekspresję genów dla widocznego regionu, czyli regionu SHC. Jednak nie tylko superskręcenie jest regulatorem tego klastra. Domniemane jest, że wpływ na niego mają również inne czynniki. W genomie *S. coelicolor* występuje system dwuskładnikowy o nazwie SatKR, składający się z genu *satR* (SCO3389) i *satK* (SCO3390) [69]. Geny te zostały odnalezione jako ważne dla badań SHC poprzez losową mutację transpozonową. Szczepy użyte w badaniach to M145 wymieniony wcześniej, jak i szczepy z mutacjami dotyczącymi SatKR. Mutanty podzielone ze względu na wprowadzone w nich mutacje SatKR. MGM11, czyli mutant o podwyższonej ekspresji *satR*, MGM12, z delecją całego systemu SatKR i MGM14 ze zwiększoną ekspresją *satR* i delecją genu *satK*.



Rysunek 4.15 Wykres powstały przy użyciu aplikacji dla dwóch publikacji opierających się na badaniu regionu SHC u *S. coelicolor*.

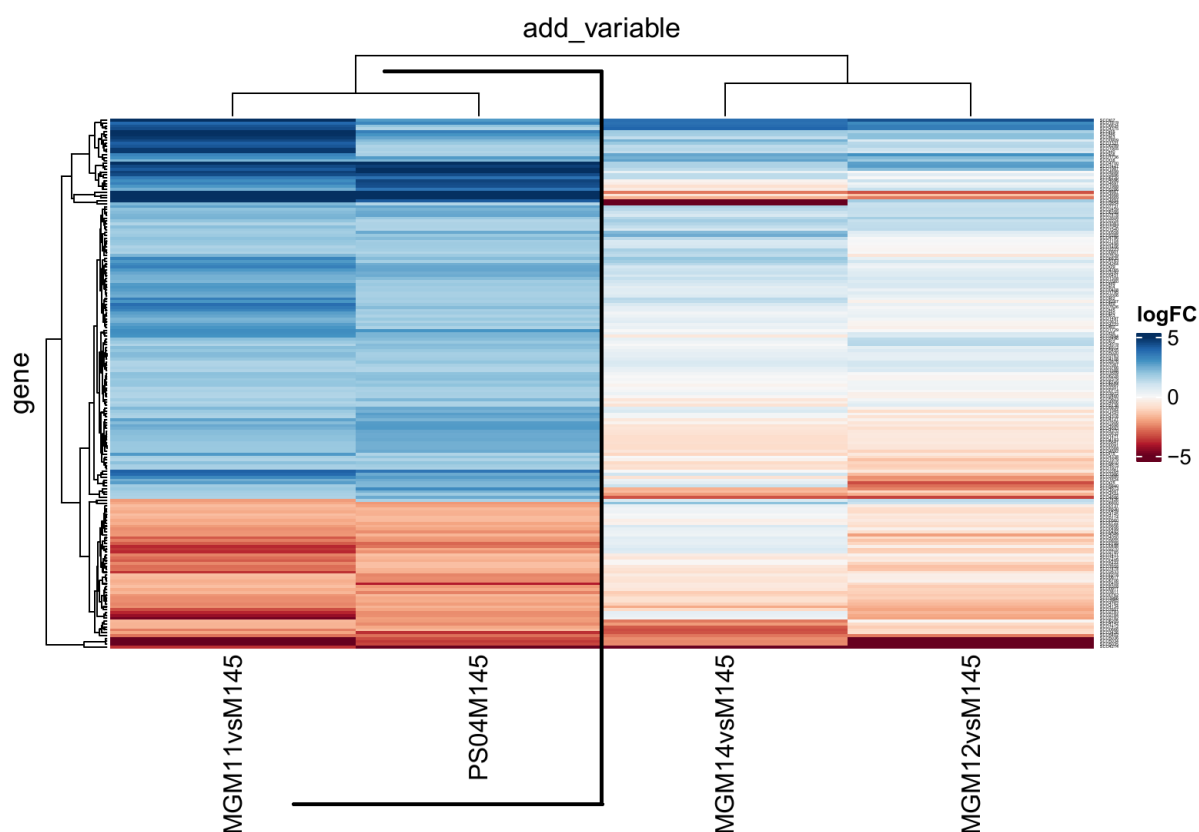
Pierwszy wniosek, który można wyciągnąć to, że największe różnice zauważalne są w porównaniach między typem dzikim, a mutantami SatKR. Samo zwiększenie ekspresji *satR* doprowadziło do wzrostu ekspresji genów klastra SHC, natomiast kombinacje delecji obu genów systemu SatKR i zwiększenia ekspresji *satR* z delecją samego *satK*, doprowadziło do znacznego obniżenia ekspresji genów klastra SHC. Można zatem zauważyć, że wynik dla szczepu dzikiego porównanego z samą deplecją TopA, z poprzedniego eksperymentu, prezentuje się podobnie do wyniku dla porównania typu dzikiego z mutantem o zwiększonej ekspresji *satR*, pomimo braku zwiększenia superskręceń DNA w tym szczepie. Stwierdzić zatem można, że system SatKR posiada właściwości regulujące klaster genów SHC. Zauważyć można, że szczepy, które posiadają delecje genu *satK* prezentują się porównywalnie, co oznaczałoby, że nie ważna jest ekspresja samego *satR*, a kluczowym komponentem dla jego działania jest drugi gen w tym systemie. Szczepy posiadające delecję genów systemu SatKR mają zauważalnie niski poziom ekspresji genów regionu SHC. Sugeruje to, że nawet w warunkach normalnego funkcjonowania komórki pod kątem superskręceń i TopA, system ten ma duże znaczenie w regulacji klastra SHC. Te wnioski sugerują, że badany system musi odgrywać kluczową rolę w regulacji genów zawartych w klastrze SHC.

Aby przeanalizować geny badanych szczepów pod szerszym kątem można zastosować inne wykresy dostępne w programie, a dokładniej diagram Venna i Heatmapę. Do analizy zostały wybrane te same cztery porównania z danych eksperymentów badających SHC. Na podstawie samej analizy RNA-seq, zauważyć można było, że w rejonie SHC, podobną ekspresję genów w porównaniu z typem dzikim, charakteryzują się szczepy PS04 i MGM11. Po wykonaniu diagramu Venna, największą liczbę genów wspólnych i istotnych statystycznie, zawierają właśnie te dwie grupy [Rys. 4.16].



Rysunek 4.16 Diagram Venna powstały przy użyciu aplikacji dla dwóch publikacji opierających się na badaniu regionu SHC u *S. coelicolor*. Zaznaczona największa część wspólna między dwoma zbiorami.

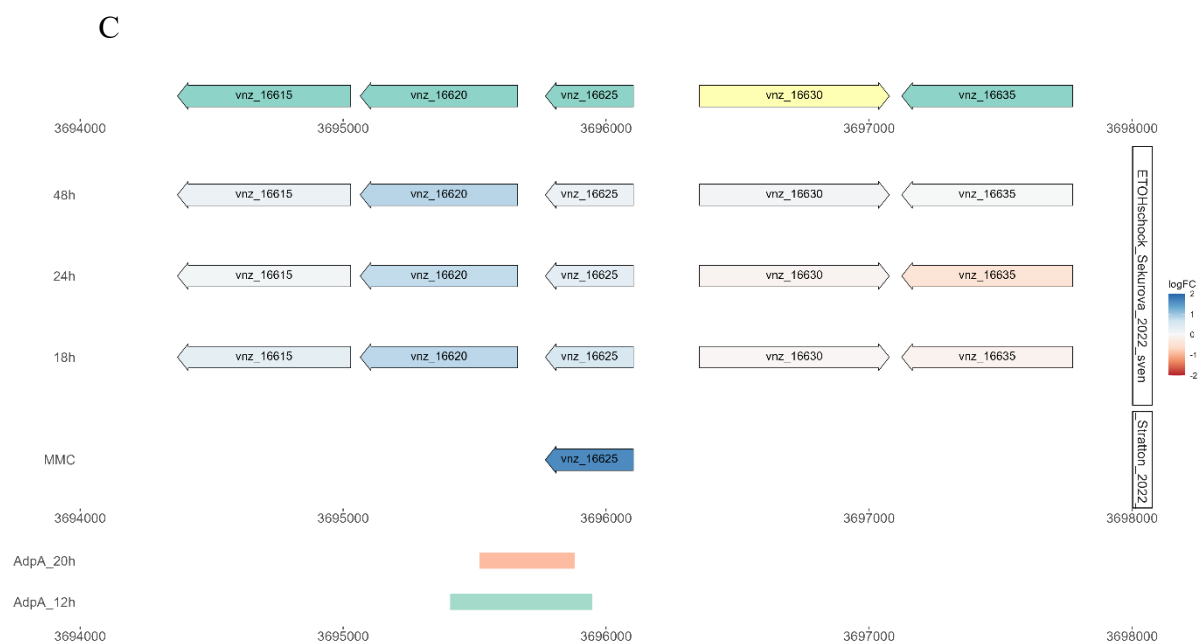
Następnym krokiem było skopiowanie wszystkich genów wspólnych dla tych dwóch zbiorów i wprowadzenie ich do okienka Heatmapy, w ten sposób można pokazać więcej genów niż tylko sam klaster SHC. Na podstawie powstałej Heatmapy, [Rys. 4.17], dojść można do wniosku, że podobieństwa pomiędzy szczepem PS04 i MGM11 nie ograniczają się tylko do regionu SHC, ale obejmują też geny rozmieszczone wzdłuż całego chromosomu *S. coelicolor*. Taki wynik może sugerować, że zarówno obniżony poziom topoizomerazy, jaki i podwyższona ekspresja genu *satR* ma podobny wpływ na regulację znacznej części z tych genów.



Rysunek 4.17 Heatmapa powstała przy użyciu aplikacji dla dwóch publikacji opierających się na badaniu regionu SHC u *S. coelicolor*. Zaznaczone zostały geny dla dwóch zbiorów, określonych jako podobne pod kątem ekspresji genów.

4.2.3 Poznanie funkcji nieznanego białka na podstawie zmian ekspresji

W wielu prokariota można odnaleźć homologi różnego rodzaju białek. Dzięki poszukiwaniu homologów dąży się do zrozumienia funkcji genów i białek między gatunkami, jak i poszerzeniu wiedzy o ich ewolucji. Jednym z interesujących genów w przypadku *S. venezuelae* jest *vnz_16625*. Na chwilę obecną podobieństwo sekwencji i struktury trzeciorzędowej białka, wskazuje, że może być on homologiem białka YbaB występującego u takich gatunków jak *E. coli*, czy *Borrelia burgdorferi*. Białko to pełni wiele funkcji, ale między innymi przypisuje mu się właściwości pozwalające na regulację odpowiedzi na stres komórki. Gen kodujący YbaB najczęściej występuje w sąsiedztwie genu *recR*. W przypadku *S. venezuelae* za gen *recR* uznaje się *vnz_16620*. Ponieważ AdpA jest globalnym regulatorem, który między innymi odpowiada za inicjowanie transkrypcji genów w odpowiedzi na czynniki środowiskowe, został dodany wykres ChIP-seq, w celu sprawdzenia, czy te geny są przez niego regulowane [Rys. 4.18].



Rysunek 4.18 Wykres powstały przy użyciu aplikacji dla dwóch publikacji związanych z wywołaniem stresu na komórce poprzez podanie czynników degradujących DNA.

Zauważyć można, że pomiędzy domniemywanym homologiem YbaB, a *recR*, znajduje się miejsce wiązania właśnie tego globalnego regulatora. Wygenerowany został wykres, dla dwóch eksperymentów, gdzie komórki *S. venezuelae* były poddawane stresowi etanolowemu i stresowi wywołanemu poprzez dodawanie mitobiocyny C, czyli tak jak etanol, czynnika degradującego DNA w komórce [Rys. 4.18]. Na podstawie zaprezentowanych wyników można zauważyć, że w przypadku szoku etanolowego ekspresja genu *vnz_16625* jest lekko podwyższona w stosunku do typu dzikiego bakterii. W przypadku traktowania komórek mitomycyną C, zauważyć można zdecydowane podwyższenie ekspresji tego genu ($\logFC = 1,63$). Wnioski te, mogą świadczyć o udziale białka Vnz_16625 w procesach naprawczych DNA w odpowiedzi na stres, lecz jest to tematyka do obszerniejszego zbadania.

5. Dyskusja

Celem pracy było stworzenie bioinformatycznego narzędzia, które będzie gromadziło wiele wyników dostępnych w publikacjach na temat *S. coelicolor* i *S. venezuelae*, jednocześnie udostępniając je w jednym miejscu w celu szybkiej i prostej analizy. Oprócz tego zależało nam, aby funkcją dostępną w programie było analizowanie własnych danych. Program miał zapewniać generowanie wykresów, które pozwalają na porównawczą analizę różnych danych, a do tego opcji pobrania powstałych wykresów. Te cele udało się osiągnąć. Powstała aplikacja spełnia wszystkie założenia, które zostały postawione na początku, choć w trakcie pracy powstały nowe pomysły, których nie udało się już zrealizować.

W „Strep.N.A” znaleźć można kilka rodzajów wykresów służących do wyświetlania danych w niej zawartych jak i załadowanych przez użytkownika. Przede wszystkim dane które odnajdziemy w aplikacji, i które możemy załadować to wyniki z badań RNA-seq, mikromacierzy i ChIP-seq, *S. coelicolor* i *S. venezuelae*. Są one uporządkowane, podzielone pod względem gatunku i dostępne dla użytkownika od razu po włączeniu aplikacji. Dwie główne strony aplikacji, można podzielić pod względem szerokości analizy jaką chcemy wykonać. Jeśli zależy nam bardziej na przeanalizowaniu niedużego fragmentu genomu takiego jak konkretny klaster genów biosyntezy jakiegoś związku, wtedy należy skorzystać ze strony pierwszej aplikacji. Dostępne są tam dwa typy wykresów, jeden dla RNA-seq i mikromacierzy, a drugi dla ChIP-seq. Poprzez wybranie różnych danych i zestawienia ich ze sobą na wykresach, w szybki i prosty sposób można wyciągnąć pierwsze wnioski dotyczące swoich badań. Podobnie jak w przykładach działania aplikacji zaprezentowanych w sekcji „Wyniki” tej pracy, po odpowiednim ustawieniu programu można zobaczyć zależności między konkretnymi genami, a jednostkami je regulującymi jak w przypadku badania AdpA i chloramfenikolu, czy zależność między różnymi mutacjami w szczepach bakterii i jak one wpływają na ekspresję niektórych genów, jak w przypadku badań dotyczących regionu SHC [68]. Natomiast gdy chcemy dokonać szerszego porównania poziomu ekspresji genów między dwoma eksperymentami należy wybrać drugą stronę aplikacji. Dzięki zastosowaniu diagramu Venna bądź wykresu Upsetplot, szybko otrzymamy wiedzę na temat podobieństw między dwoma eksperymentami. Następnie z uzyskanych przy użyciu diagramu Venna, części wspólnych można stworzyć Heatmapę. W ten sposób otrzymać można obszerniejsze spojrzenie na geny istotne w danych badaniach oraz jak one prezentują się między sobą. Można zaobserwować takie zależności jak bardziej globalne podobieństwo między ekspresją szczepów, podobnie jak w przypadku badań dotyczących SHC.

W temacie zastosowania aplikacji pozostając, należy podkreślić, że nie jest to narzędzie, które wykonuje obliczenia dla surowych wyników z eksperymentów. Jest to narzędzie służące do analizy już wstępnie znormalizowanych i przeliczonych wyników. Program nie jest w stanie sam obliczyć wartości parametrów takich jak logFC, czy FDR z czystych odczytów, operuje on tylko na gotowych wynikach. W praktyce oznacza to, że program jest przydatny na końcowym etapie analiz transkryptomicznych. Aby w pełni skorzystać z programu, można oprzeć się na obecnych w niej danych, ale gdy chcemy porównać je ze swoimi, musimy mieć już gotowy plik zawierający wartości logFC i FDR.

Dzięki zastosowaniu aplikacji, możemy sobie odpowiedzieć na takie pytania jak: czy w całym genomie można zauważyć odstępstwa w ekspresji, czy konkretny klastery genów ma znacząco zmienioną ekspresję, czy interesujący nas regulator przyłącza się w przypuszczanym miejscu, i co to może oznaczać. Istnieje wiele programów, które pozwalają na podobne zestawienia i analizy porównawcze, ale posiadają one jeden podstawowy problem. Mianowicie nie dotyczą one *S. coelicolor* i *S. venezuelae*. W programie znajduje się znacznie mniej danych niż w najpopularniejszych bazach danych dotyczących innych organizmów, jednak jest jedynym, skupionym na tych dwóch gatunkach *Streptomyces* pod kątem badań transkryptomicznych. Nie posiada on mocy obliczeniowej takiej jak popularne strony do analiz np. Galaxy, czy iDEP, lecz to nie było naszym zamiarem. Chcieliśmy stworzyć interfejs, który umożliwi szybką analizę porównawczą, w sposób graficzny, a nie narzędzie obliczeniowe, ponieważ, między innymi, wymienione wcześniej dwie platformy są popularnym rozwiązaniem do tego służącym. Innym podobnym narzędziem stosowanym do analiz, który częściowo spełnia kryteria tego co chcieliśmy osiągnąć w naszym programie jest GeneTonic [89]. Przy użyciu tego pakietu możemy wykonać wiele analiz danych transkryptomicznych, aczkolwiek potrzebna jest chociaż podstawa wiedzy tego czym jest język programowania R. Nawet jeśli z tym przeciętny użytkownik prawdopodobnie sobie poradzi, to „Strep.N.A” ponownie przeważa nad pakietem GeneTonic tym, że zawiera przygotowaną bazę danych, a GeneTonic, takiej nie posiada.

Problemem, z którym należało się zmierzyć podczas powstawania programu, był format danych oraz ich poprawne załadowanie. Na początku zauważono, że zgromadzone pliki znacznie różniły się między sobą – zarówno pod względem rozszerzeń („.xlsx”, „.txt”, PDF), jak i układu kolumn oraz sposobu prezentacji wyników. W związku z tym opracowaliśmy własny, ujednolicony schemat zapisu, który zawiera wszystkie kluczowe parametry potrzebne do analizy, takie jak logFC, FDR czy nazwy genów. Jednocześnie

zachowano oryginalne wartości wyników, z minimalną ingerencją w dane źródłowe. Dodatkowym utrudnieniem było to, że część badaczy udostępniała jedynie statystycznie istotne wyniki w publikacji, pomijając pełne zestawienia w suplementach. Z tego powodu niektóre dane w aplikacji nie zawierają kompletnego obrazu analizy transkryptomicznej dla danego badania. Kolejnym aspektem była kwestia normalizacji danych. Różnorodność metod i programów stosowanych przez autorów badań sprawia, że nie jesteśmy w stanie zagwarantować jednolitego podejścia przetworzenia wyników wyjściowych.

Wspomniany ujednolicony układ danych przygotowywano głównie przy użyciu Excela oraz języka R. Ze względu na znaczne różnice w strukturze plików, jakie występują w analizach transkryptomicznych, sam arkusz kalkulacyjny okazywał się często niewystarczający. Mając świadomość, że nie każdy użytkownik posługuje się R, zaproponowaliśmy dodatkowy, prostszy sposób przygotowania danych możliwy do wykonania w Excelu. Tak zapisane pliki są następnie automatycznie przekształcane w programie z użyciem skryptów w R.

Przyszłość naszego programu opierała będzie się przede wszystkim na rozszerzaniu jego bazy danych o nowsze eksperymenty, aby pozostawał użytkowy. Już na etapie końca pracy nad aplikacją, wiadomym było, że baza danych którą udało nam się zgromadzić jest ograniczona i warto ją rozszerzyć, niestety na to nie starczyło czasu. Nie oznacza to jednocześnie, że rozmiar bazy jest nie wystarczający do przeprowadzania analiz, po prostu nie znajdziemy tam każdego przypadku dotyczącego *Streptomyces*. Oprócz dodawania nowych danych, istotne również jest poprawienie programu pod względem optymalizacji jego działania. Nawiązując do wpływu ładowania danych własnych, na obecne już załadowane, jedną z funkcji, którą można wprowadzić, aby usprawnić ten proces jest asynchroniczne ładowanie danych. Asynchroniczne podejście do ładowania danych zapewni również, wczytywanie ich tylko gdy będą wymagane, w odróżnieniu od aktualnego działania aplikacji polegającego na wczytaniu wszystkich danych przy jej uruchamianiu. Będzie to kluczowa zmiana w momencie, gdy danych w bazie znajdzie się dużo więcej. Poza poprawianiem istniejących fragmentów aplikacji, można dodać kilka nowych opcji. Jak wspomniane wcześniej, podczas powstawania programu powstało wiele pomysłów na inne typy wykresów czy możliwości obliczeniowe, które można by było wprowadzić. Między innymi podjęta zostanie próba zmienienia statycznych wykresów na ich interaktywne wersje, gdzie najechnie myszką na gen wyświetli na przykład podstawowe informacje o nim bądź będzie można wybrać poszczególne geny w celu zaznaczenia ich innym kolorem. Celem na

przyszłość jest również wprowadzenie dwóch innych wykresów do aplikacji. Jednym z nich jest wykres przedstawiający znormalizowany poziom ekspresji genów (RPKM, Reads Per Kilobase of transcript per Million mapped reads), często stosowany w analizach danych transkryptomicznych, oraz wykres pokazujący zmianę ekspresji genów w czasie. Niestety, na chwilę obecną nie było możliwości wprowadzenia tych funkcji do programu. Operowałyby one na danych przedstawiających bazowe odczyty dla poszczególnych genów, których nie udało się nam zgromadzić w wystarczającej ilości.

Program „Strep.N.A” ma na celu ułatwienie użytkownikom analizy swoich wyników danych transkryptomicznych *S. coelicolor* i *S. venezuelae*. Jak zostało zaprezentowane w tej pracy, jest to program łatwy w użyciu, przystępny dla użytkownika i przy jego zastosowaniu można szybko wyciągnąć wnioski na temat swoich danych bądź tych obecnych w bazie programu. Wersja udostępniana na chwilę obecną jest wersją wstępną i mamy nadzieję na pomyślne rozszerzanie programu po zakończeniu pracy nad nim w ramach projektu magisterskiego.

6. Bibliografia

- [1] E. A. Barka et al., “Taxonomy, Physiology, and Natural Products of Actinobacteria,” *Microbiology and Molecular Biology Reviews*, vol. 80, no. 1, pp. 1–43, Mar. 2016, doi: 10.1128/MMBR.00019-15.
- [2] S. D. Bentley et al., “Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2),” *Nature*, vol. 417, no. 6885, pp. 141–147, May 2002, doi: 10.1038/417141a.
- [3] M. Nikolaidis et al., “A panoramic view of the genomic landscape of the genus *Streptomyces*,” *Microb Genom*, vol. 9, no. 6, Jun. 2023, doi: 10.1099/mgen.0.001028.
- [4] S. E. Jones and M. A. Elliot, “‘Exploring’ the regulation of *Streptomyces* growth and development,” *Curr Opin Microbiol*, vol. 42, pp. 25–30, Apr. 2018, doi: 10.1016/j.mib.2017.09.009.
- [5] G. A. Quinn, A. M. Banat, A. M. Abdelhameed, and I. M. Banat, “*Streptomyces* from traditional medicine: sources of new innovations in antibiotic discovery,” *J Med Microbiol*, vol. 69, no. 8, pp. 1040–1048, Aug. 2020, doi: 10.1099/jmm.0.001232.
- [6] K. F. Chater, “Recent advances in understanding *Streptomyces*,” *F1000Res*, vol. 5, p. 2795, Nov. 2016, doi: 10.12688/f1000research.9534.1.
- [7] D. L. Sexton and E. I. Tocheva, “Ultrastructure of Exospore Formation in *Streptomyces* Revealed by Cryo-Electron Tomography,” *Front Microbiol*, vol. 11, Sep. 2020, doi: 10.3389/fmicb.2020.581135.
- [8] K. Alam et al., “*Streptomyces*: The biofactory of secondary metabolites,” *Front Microbiol*, vol. 13, Sep. 2022, doi: 10.3389/fmicb.2022.968053.
- [9] Z. Liu, Y. Zhao, C. Huang, and Y. Luo, “Recent Advances in Silent Gene Cluster Activation in *Streptomyces*,” *Front Bioeng Biotechnol*, vol. 9, Feb. 2021, doi: 10.3389/fbioe.2021.632230.
- [10] S. Hwang et al., “*Streptomyces* as Microbial Chassis for Heterologous Protein Expression,” *Front Bioeng Biotechnol*, vol. 9, Dec. 2021, doi: 10.3389/fbioe.2021.804295.

- [11] S. Schlimpert and M. A. Elliot, “The Best of Both Worlds—*Streptomyces coelicolor* and *Streptomyces venezuelae* as Model Species for Studying Antibiotic Production and Bacterial Multicellular Development,” *J Bacteriol*, vol. 205, no. 7, Jul. 2023, doi: 10.1128/jb.00153-23.
- [12] L. V Bystrykh, M. A. Fernández-Moreno, J. K. Herrema, F. Malpartida, D. A. Hopwood, and L. Dijkhuizen, “Production of actinorhodin-related ‘blue pigments’ by *Streptomyces coelicolor* A3(2),” *J Bacteriol*, vol. 178, no. 8, pp. 2238–2244, Apr. 1996, doi: 10.1128/jb.178.8.2238-2244.1996.
- [13] D. Sun, C. Liu, J. Zhu, and W. Liu, “Connecting Metabolic Pathways: Sigma Factors in *Streptomyces* spp.,” *Front Microbiol*, vol. 8, Dec. 2017, doi: 10.3389/fmicb.2017.02546.
- [14] E. Lee *et al.*, “A master regulator σ B governs osmotic and oxidative response as well as differentiation via a network of sigma factors in *Streptomyces coelicolor*,” *Mol Microbiol*, vol. 57, no. 5, pp. 1252–1264, Sep. 2005, doi: 10.1111/j.1365-2958.2005.04761.x.
- [15] A. Romero-Rodríguez, I. Robledo-Casados, and S. Sánchez, “An overview on transcriptional regulators in *Streptomyces*,” *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1849, no. 8, pp. 1017–1039, Aug. 2015, doi: 10.1016/j.bbagrm.2015.06.007.
- [16] M. Płachetka, M. Krawiec, J. Zakrzewska-Czerwińska, and M. Wolański, “AdpA Positively Regulates Morphological Differentiation and Chloramphenicol Biosynthesis in *Streptomyces venezuelae*,” *Microbiol Spectr*, vol. 9, no. 3, Dec. 2021, doi: 10.1128/Spectrum.01981-21.
- [17] M. J. Szafran, D. Jakimowicz, and M. A. Elliot, “Compaction and control—the role of chromosome-organizing proteins in *Streptomyces*,” *FEMS Microbiol Rev*, vol. 44, no. 6, pp. 725–739, Nov. 2020, doi: 10.1093/femsre/fuaa028.
- [18] A. Strzałka, J. Mikołajczyk, K. Kowalska, M. Skurczyński, N. A. Holmes, and D. Jakimowicz, “The role of two major nucleoid-associated proteins in *Streptomyces*, HupA and HupS, in stress survival and gene expression regulation,” *Microb Cell Fact*, vol. 23, no. 1, p. 275, Oct. 2024, doi: 10.1186/s12934-024-02549-0.
- [19] X. Xia, J. Zhang, J. Zheng, G. Liao, Y. Ding, and Y. Li, “Important Role of Bacterial Nucleoid-Associated Proteins in Discovery of Novel Secondary Metabolites,” *Int J Mol Sci*, vol. 26, no. 6, p. 2393, Mar. 2025, doi: 10.3390/ijms26062393.

- [20] E. M. F. Shepherdson and M. A. Elliot, “Redefining development in *Streptomyces venezuelae* : integrating exploration into the classical sporulating life cycle,” *mBio*, vol. 15, no. 4, Apr. 2024, doi: 10.1128/mbio.02424-23.
- [21] N. Tschowri, “Cyclic Dinucleotide-Controlled Regulatory Pathways in *Streptomyces* Species,” *J Bacteriol*, vol. 198, no. 1, pp. 47–54, Jan. 2016, doi: 10.1128/JB.00423-15.
- [22] L. T. Fernández-Martínez *et al.*, “New Insights into Chloramphenicol Biosynthesis in *Streptomyces venezuelae* ATCC 10712,” *Antimicrob Agents Chemother*, vol. 58, no. 12, pp. 7441–7450, Dec. 2014, doi: 10.1128/AAC.04272-14.
- [23] D. Muzzey, E. A. Evans, and C. Lieber, “Understanding the Basics of NGS: From Mechanism to Variant Calling,” *Curr Genet Med Rep*, vol. 3, no. 4, pp. 158–165, Dec. 2015, doi: 10.1007/s40142-015-0076-8.
- [24] M. Marcinkowska-Swojak, M. Rakoczy, J. Podkowiński, J. Handschuh, P. Wojciechowski, and L. Handschuh, “Od Sangera do sekwencjonowania genomów – przegląd technologii sekwencjonowania DNA,” *Postepy Biochem*, vol. 70, no. 2, pp. 173–189, Jul. 2024, doi: 10.18388/pb.2021_534.
- [25] S. Marguerat and J. Bähler, “RNA-seq: from technology to biology,” *Cellular and Molecular Life Sciences*, vol. 67, no. 4, pp. 569–579, Feb. 2010, doi: 10.1007/s00018-009-0180-6.
- [26] K. Van Den Berge *et al.*, “RNA sequencing data: hitchhiker’s guide to expression analysis,” Oct. 17, 2018. doi: 10.7287/peerj.preprints.27283v1.
- [27] S. Antoraz *et al.*, “The Orphan Response Regulator Aor1 Is a New Relevant Piece in the Complex Puzzle of *Streptomyces coelicolor* Antibiotic Regulatory Network,” *Front Microbiol*, vol. 8, Dec. 2017, doi: 10.3389/fmicb.2017.02444.
- [28] Y. Lu *et al.*, “An Orphan Histidine Kinase, OhkA, Regulates Both Secondary Metabolism and Morphological Differentiation in *Streptomyces coelicolor*,” *J Bacteriol*, vol. 193, no. 12, pp. 3020–3032, Jun. 2011, doi: 10.1128/JB.00017-11.
- [29] Y. Jeong *et al.*, “The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2),” *Nat Commun*, vol. 7, no. 1, p. 11605, Jun. 2016, doi: 10.1038/ncomms11605.
- [30] Joanne. Jones, “The Basics of Microarrays,” *ACNR*, p. 29, 2006.

- [31] P. Salerno *et al.*, “Identification of new developmentally regulated genes involved in *Streptomyces coelicolor* sporulation,” *BMC Microbiol*, vol. 13, no. 1, p. 281, Dec. 2013, doi: 10.1186/1471-2180-13-281.
- [32] P. Yagüe *et al.*, “Transcriptomic Analysis of *Streptomyces coelicolor* Differentiation in Solid Sporulating Cultures: First Compartmentalized and Second Multinucleated Mycelia Have Different and Distinctive Transcriptomes,” *PLoS One*, vol. 8, no. 3, p. e60665, Mar. 2013, doi: 10.1371/journal.pone.0060665.
- [33] User:Larssono, “No Title.” [Online]. Available: <https://commons.wikimedia.org/wiki/File:Microarray-schema.jpg>
- [34] P. J. Park, “ChIP-seq: advantages and challenges of a maturing technology,” *Nat Rev Genet*, vol. 10, no. 10, pp. 669–680, Oct. 2009, doi: 10.1038/nrg2641.
- [35] M. J. Szafran *et al.*, “Spatial rearrangement of the *Streptomyces venezuelae* linear chromosome during sporogenic development,” *Nat Commun*, vol. 12, no. 1, p. 5222, Sep. 2021, doi: 10.1038/s41467-021-25461-2.
- [36] K. Šmídová, A. Ziková, J. Pospíšil, M. Schwarz, J. Bobek, and J. Vohradsky, “DNA mapping and kinetic modeling of the HrdB regulon in *Streptomyces coelicolor*,” *Nucleic Acids Res*, vol. 47, no. 2, pp. 621–633, Jan. 2019, doi: 10.1093/nar/gky1018.
- [37] A. Shah, “Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system,” *Nat Methods*, vol. 6, no. 4, pp. ii–iii, Apr. 2009, doi: 10.1038/nmeth.f.247.
- [38] C. Gao, Hindra, D. Mulder, C. Yin, and M. A. Elliot, “Crp Is a Global Regulator of Antibiotic Production in *Streptomyces*,” *mBio*, vol. 3, no. 6, Dec. 2012, doi: 10.1128/mBio.00407-12.
- [39] R. Edgar, “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository,” *Nucleic Acids Res*, vol. 30, no. 1, pp. 207–210, Jan. 2002, doi: 10.1093/nar/30.1.207.
- [40] U. Sarkans *et al.*, “The BioStudies database—one stop shop for all data supporting a life sciences study,” *Nucleic Acids Res*, vol. 46, no. D1, pp. D1266–D1270, Jan. 2018, doi: 10.1093/nar/gkx965.
- [41] “The Cancer Genome Atlas.” [Online]. Available: <https://www.cancer.gov/tcga>

- [42] K. Rychel, K. Decker, A. V Sastry, P. V Phaneuf, S. Poudel, and B. O. Palsson, “iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning,” *Nucleic Acids Res*, vol. 49, no. D1, pp. D112–D120, Jan. 2021, doi: 10.1093/nar/gkaa810.
- [43] L. A. L. Abueg *et al.*, “The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update,” *Nucleic Acids Res*, vol. 52, no. W1, pp. W83–W94, Jul. 2024, doi: 10.1093/nar/gkae410.
- [44] J. T. Robinson *et al.*, “Integrative genomics viewer,” *Nat Biotechnol*, vol. 29, no. 1, pp. 24–26, Jan. 2011, doi: 10.1038/nbt.1754.
- [45] J. T. Robinson *et al.*, “Integrative genomics viewer,” *Nat Biotechnol*, vol. 29, no. 1, pp. 24–26, Jan. 2011, doi: 10.1038/nbt.1754.
- [46] S. X. Ge, E. W. Son, and R. Yao, “iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data,” *BMC Bioinformatics*, vol. 19, no. 1, p. 534, Dec. 2018, doi: 10.1186/s12859-018-2486-6.
- [47] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov, “GenePattern 2.0,” *Nat Genet*, vol. 38, no. 5, pp. 500–501, May 2006, doi: 10.1038/ng0506-500.
- [48] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, “ToppGene Suite for gene list enrichment analysis and candidate gene prioritization,” *Nucleic Acids Res*, vol. 37, no. Web Server, pp. W305–W311, Jul. 2009, doi: 10.1093/nar/gkp427.
- [49] S. X. Ge, D. Jung, and R. Yao, “ShinyGO: a graphical gene-set enrichment tool for animals and plants,” *Bioinformatics*, vol. 36, no. 8, pp. 2628–2629, Apr. 2020, doi: 10.1093/bioinformatics/btz931.
- [50] R Core Team, “R: A Language and Environment for Statistical Computing,” 2023, *Vienna, Austria*. [Online]. Available: <https://www.r-project.org/>
- [51] RStudio Team, “RStudio: Integrated Development Environment for R,” 2020, *Boston, MA*. [Online]. Available: <http://www.rstudio.com/>
- [52] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: <https://ggplot2.tidyverse.org>
- [53] D. Wilkins, “gggenes: Draw Gene Arrow Maps in ‘ggplot2,’” 2023. [Online]. Available: <https://cran.r-project.org/package=gggenes>

- [54] W. Chang *et al.*, “shiny: Web Application Framework for R,” 2023. [Online]. Available: <https://cran.r-project.org/package=shiny>
- [55] H. Wickham, “Tidy Data,” *J Stat Softw*, vol. 59, no. 10, 2014, doi: 10.18637/jss.v059.i10.
- [56] W. Chang, “shinythemes: Themes for Shiny,” 2021. [Online]. Available: <https://cran.r-project.org/package=shinythemes>
- [57] H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan, “dplyr: A Grammar of Data Manipulation,” 2023. [Online]. Available: <https://cran.r-project.org/package=dplyr>
- [58] H. Wickham, D. Vaughan, and M. Girlich, “tidyr: Tidy Messy Data,” 2023. [Online]. Available: <https://cran.r-project.org/package=tidyr>
- [59] T. L. Pedersen, “patchwork: The Composer of Plots,” 2024. [Online]. Available: <https://cran.r-project.org/package=patchwork>
- [60] M. S. and P. A., “tidyHeatmap: an R package for modular heatmap production based on tidy principles,” *J Open Source Softw*, vol. 5, no. 52, p. 2472, 2020, doi: 10.21105/joss.02472.
- [61] M. Morgan and M. Ramos, “BiocManager: Access the Bioconductor Project Package Repository,” 2023. [Online]. Available: <https://cran.r-project.org/package=BiocManager>
- [62] H. Wickham, “stringr: Simple, Consistent Wrappers for Common String Operations,” 2023. [Online]. Available: <https://cran.r-project.org/package=stringr>
- [63] C. Sievert, J. Cheng, and G. Aden-Buie, “bslib: Custom ‘Bootstrap’ ‘Sass’ Themes for ‘shiny’ and ‘rmarkdown,’” 2023. [Online]. Available: <https://cran.r-project.org/package=bslib>
- [64] L. Yan, “ggvenn: Draw Venn Diagram by ‘ggplot2,’” 2023. [Online]. Available: <https://cran.r-project.org/package=ggvenn>
- [65] C. Ahlmann-Eltze, “ggupset: Combination Matrix Axis for ‘ggplot2’ to Create ‘UpSet’ Plots,” 2024. [Online]. Available: <https://cran.r-project.org/package=ggupset>
- [66] J. Ooms, “pdftools: Text Extraction, Rendering and Converting of PDF Documents,” 2024. [Online]. Available: <https://cran.r-project.org/package=pdfutils>
- [67] S. Urbanek, “png: Read and write PNG images,” 2022. [Online]. Available: <https://cran.r-project.org/package=png>

- [68] M. J. Szafran, M. Gongerowska, T. Małecki, M. Elliot, and D. Jakimowicz, “Transcriptional Response of *Streptomyces coelicolor* to Rapid Chromosome Relaxation or Long-Term Supercoiling Imbalance,” *Front Microbiol*, vol. 10, Jul. 2019, doi: 10.3389/fmicb.2019.01605.
- [69] M. Gongerowska-Jac *et al.*, “Global Chromosome Topology and the Two-Component Systems in Concerted Manner Regulate Transcription in *Streptomyces*,” *mSystems*, vol. 6, no. 6, Dec. 2021, doi: 10.1128/mSystems.01142-21.
- [70] R. Sánchez de la Nieta, S. Antoraz, J. F. Alzate, R. I. Santamaría, and M. Díaz, “Antibiotic Production and Antibiotic Resistance: The Two Sides of AbrB1/B2, a Two-Component System of *Streptomyces coelicolor*,” *Front Microbiol*, vol. 11, Oct. 2020, doi: 10.3389/fmicb.2020.587750.
- [71] S. Rico *et al.*, “Deciphering the Regulon of *Streptomyces coelicolor* AbrC3, a Positive Response Regulator of Antibiotic Production,” *Appl Environ Microbiol*, vol. 80, no. 8, pp. 2417–2428, Apr. 2014, doi: 10.1128/AEM.03378-13.
- [72] A. Botas *et al.*, “ArgR of *Streptomyces coelicolor* Is a Pleiotropic Transcriptional Regulator: Effect on the Transcriptome, Antibiotic Production, and Differentiation in Liquid Cultures,” *Front Microbiol*, vol. 9, Mar. 2018, doi: 10.3389/fmicb.2018.00361.
- [73] C. D. Den Hengst, N. T. Tran, M. J. Bibb, G. Chandra, B. K. Leskiw, and M. J. Buttner, “Genes essential for morphological development and antibiotic production in *Streptomyces coelicolor* are targets of BldD during vegetative growth,” *Mol Microbiol*, vol. 78, no. 2, pp. 361–379, Oct. 2010, doi: 10.1111/j.1365-2958.2010.07338.x.
- [74] M. J. Bush, G. Chandra, M. M. Al-Bassam, K. C. Findlay, and M. J. Buttner, “BldC Delays Entry into Development To Produce a Sustained Period of Vegetative Growth in *Streptomyces venezuelae*,” *mBio*, vol. 10, no. 1, Feb. 2019, doi: 10.1128/mBio.02812-18.
- [75] Z. Yu, H. Zhu, G. Zheng, W. Jiang, and Y. Lu, “A genome-wide transcriptomic analysis reveals diverse roles of the two-component system DraR-K in the physiological and morphological differentiation of *Streptomyces coelicolor*,” *Appl Microbiol Biotechnol*, vol. 98, no. 22, pp. 9351–9363, Nov. 2014, doi: 10.1007/s00253-014-6102-z.
- [76] Q. Liu, D. Pinto, and T. Mascher, “Characterization of the Widely Distributed Novel ECF42 Group of Extracytoplasmic Function σ Factors in *Streptomyces venezuelae*,” *J Bacteriol*, vol. 200, no. 21, Nov. 2018, doi: 10.1128/JB.00437-18.

- [77] S. T. Pullan, G. Chandra, M. J. Bibb, and M. Merrick, “Genome-wide analysis of the role of GlnR in *Streptomyces venezuelae* provides new insights into global nitrogen regulation in actinomycetes,” *BMC Genomics*, vol. 12, no. 1, p. 175, Dec. 2011, doi: 10.1186/1471-2164-12-175.
- [78] M. Urem *et al.*, “OsdR of *Streptomyces coelicolor* and the Dormancy Regulator DevR of *Mycobacterium tuberculosis* Control Overlapping Regulons,” *mSystems*, vol. 1, no. 3, Jun. 2016, doi: 10.1128/mSystems.00014-16.
- [79] D. Kallifidas, D. Thomas, P. Doughty, and M. S. B. Paget, “The σ R regulon of *Streptomyces coelicolor* A3(2) reveals a key role in protein quality control during disulphide stress,” *Microbiology (N Y)*, vol. 156, no. 6, pp. 1661–1672, Jun. 2010, doi: 10.1099/mic.0.037804-0.
- [80] N. Naseer, J. A. Shapiro, and M. Chander, “RNA-Seq Analysis Reveals a Six-Gene SoxR Regulon in *Streptomyces coelicolor*,” *PLoS One*, vol. 9, no. 8, p. e106181, Aug. 2014, doi: 10.1371/journal.pone.0106181.
- [81] O. N. Sekurova, M. Zehl, M. Predl, P. Hunyadi, T. Rattei, and S. B. Zotchev, “Targeted Metabolomics and High-Throughput RNA Sequencing-Based Transcriptomics Reveal Massive Changes in the *Streptomyces venezuelae* NRRL B-65442 Metabolism Caused by Ethanol Shock,” *Microbiol Spectr*, vol. 10, no. 6, Dec. 2022, doi: 10.1128/spectrum.03672-22.
- [82] M. M. Avramova *et al.*, “Global Effects of the Developmental Regulator BldB in *Streptomyces venezuelae*,” *J Bacteriol*, vol. 205, no. 6, Jun. 2023, doi: 10.1128/jb.00135-23.
- [83] K. J. Stratton, M. J. Bush, G. Chandra, C. E. M. Stevenson, K. C. Findlay, and S. Schlimpert, “Genome-Wide Identification of the LexA-Mediated DNA Damage Response in *Streptomyces venezuelae*,” *J Bacteriol*, vol. 204, no. 8, Aug. 2022, doi: 10.1128/jb.00108-22.
- [84] Y. Lei, H. Onaka, and S. Asamizu, “Transcriptionally induced nucleoid-associated protein-like ccr1 in combined-culture serves as a global effector of *Streptomyces* secondary metabolism,” *Front Microbiol*, vol. 15, Jul. 2024, doi: 10.3389/fmicb.2024.1422977.
- [85] C. Du, J. Willemse, A. M. Erkelens, V. J. Carrion, R. T. Dame, and G. P. van Wezel, “System-Wide Analysis of the GATC-Binding Nucleoid-Associated Protein Gbn and Its Impact on *Streptomyces* Development,” *mSystems*, vol. 7, no. 3, Jun. 2022, doi: 10.1128/msystems.00061-22.

- [86] A. Strzałka *et al.*, “Enhanced binding of an HU homologue under increased DNA supercoiling preserves chromosome organisation and sustains *Streptomyces* hyphal growth,” *Nucleic Acids Res*, vol. 50, no. 21, pp. 12202–12216, Nov. 2022, doi: 10.1093/nar/gkac1093.
- [87] M. Płachetka, M. Krawiec, J. Zakrzewska-Czerwińska, and M. Wolański, “AdpA Positively Regulates Morphological Differentiation and Chloramphenicol Biosynthesis in *Streptomyces venezuelae*,” *Microbiol Spectr*, vol. 9, no. 3, Dec. 2021, doi: 10.1128/Spectrum.01981-21.
- [88] G. C. Oong and P. Tadi, *Chloramphenicol*. 2025. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/17737966>
- [89] F. Marini, A. Ludt, J. Linke, and K. Strauch, “GeneTonic: an R/Bioconductor package for streamlining the interpretation of RNA-seq data,” *BMC Bioinformatics*, vol. 22, no. 1, p. 610, Dec. 2021, doi: 10.1186/s12859-021-04461-5.

7. Skróty i sformułowania anglojęzyczne

- *Streptomyces coelicolor* – *S. coelicolor*
- *Streptomyces venezuelae* – *S. venezuelae*
- R – język programowania.
- Klaster genów biosyntezy – biosynthesis gene cluster, BGC
- Białka zaangażowane w organizację nukleoidu - Nucleoid-associated proteins, NAP
- Sekwencjonowanie nowej generacji – New generation sequencing, NGS
- Sekwencjonowanie RNA- RNA-seq
- Fold change – FC
- Wskaźnik fałszywych odczytów – false discovery rate, FDR
- Sekwencjonowanie DNA po immunoprecypitacji chromatyny – ChIP-seq
- Białko receptorowe cyklicznego AMP – cyclic AMP receptor protein, Crp
- Gene expression omnibus – GEO
- National Center for Biotechnology Information – NCBI
- The Cancer genome Atlas – TCGA
- Integrative Genomics Viewer – IGV
- Dodatkowa zmienna – additional variable, add_variable
- Typ dziki – wild type, WT
- Region superczuły na superskręcenia DNA – supercoiling-hypersensitive region, SHC