Metodologia statystyczna badania zmian jakościowych transkryptomu roślin w aspekcie interakcji genotypu i środowiska

PODZIĘKOWANIA

Pragnę podziękować Panu Promotorowi, prof. dr. hab. Pawłowi Krajewskiemu, za pomoc, bez której ta praca by nie powstała. Niezliczone rozmowy i poświęcony czas, a także pokłady cierpliwości i życzliwość sprawiły, że proces prowadzenia badań, poszukiwania rozwiązań, obliczeń i pisania był cennym i dobrym doświadczeniem przy niezmiennie towarzyszącej mi świadomości, że w razie jakiegokolwiek problemu mogę liczyć na wsparcie.

Chciałabym również podziękować Pani Promotor Pomocniczej, dr Hannie Ćwiek-Kupczyńskiej, za wszystkie rady i rozmowy – na żywo oraz przez wideopołączenia – a także za uwagi i podpowiedzi, dzięki którym mogłam dopracować niniejszy tekst.

Dziękuję również prof. dr hab. Barbarze Naganowskiej za cenne uwagi i sugestie, które pomogły mi doszlifować rozprawę doktorską i pozwoliły stworzyć pełniejszą wersję tekstu.

Dziękuję prof. dr hab. Anetcie Kuczyńskiej, dr. hab. inż. Krzysztofowi Mikołajczakowi, prof. IGR PAN i dr. Michałowi Kempie za pomoc i pozwolenie na wykorzystanie danych, na których mogłam sprawdzić przedstawioną w pracy metodę analizy.

I would like to thank dr. Franklin Gregory, prof. IGR PAN, because accepting me into the HyperiSyn project enabled me to start my journey called doctoral studies. Regular meetings with the entire team allowed me to acclimate faster and feel that I was in the right place. I would also like to thank dr Selvakesavan Rajendran Kamalabai for patiently introducing me to the world of plants, so that I would know where the data I analyze comes from.

Dziękuję również dr inż. Monice Mokrzyckiej za wsparcie i rozmowy, a także za słowa otuchy, które motywowały mnie, kiedy miewałam chwilowe zastoje w pracy.

Na koniec chciałabym podziękować mojej rodzinie: Mamie, Tacie, Babciom, Zosi i Nowakowi, którzy z delikatnością, ale i ze zdecydowaniem, konsekwentnie dopytywali o postępy, cierpliwie słuchali, doradzali i wspierali na każdym kroku.

Spis treści

Finansowanie 4

Streszczenie 5

Abstract 6

1. Wstęp 7

2. Stan wiedzy 8

2.1. Sekwencjonowanie DNA 8

2.2. Natura i badanie zmian sekwencji nukleotydowych RNA 9

2.3. Metody badania zmienności transkryptomu 12

2.3.1. Odległości i ich zastosowania 13

2.3.2. Odległości między sekwencjami 14

2.3.3. Odległość Mash 15

3. Cel pracy 21

4. Metody 22

4.1. Przetwarzanie danych sekwencyjnych 22

4.1.1. Kontrola jakości danych 22

4.1.2. Składanie odczytów 22

4.2. Liczenie odległości Mash 23

4.3. Analiza współrzędnych głównych 23

4.4. Analiza wariancji i kowariancji 24

4.5. Porównawcza analiza ekspresji genów 26

4.5.1. Mapowanie odczytów 26

4.5.2. Analiza różnicowa ekspresji genów 26

4.6. Ontologie genów 27

5. Opis danych doświadczalnych 28

6. Wyniki 29

6.1. Analiza podobieństwa transkryptomów w doświadczeniach wieloczynnikowych 29

6.1.1. Analiza odległości wykorzystująca dane NGS 35

6.1.2. Analiza odległości wykorzystująca skafoldy 40

6.2. Analiza podobieństwa transkryptomów w aspekcie funkcji molekularnych i procesów biologicznych 50
6.3. Analiza różnicowa ekspresji genów uwzględniająca różnice transkryptomu 68
7. Dyskusja 70
SUPLEMENT 78
LITERATURA 95

Finansowanie

Narodowe Centrum Nauki, Polska, projekt: Harmonia 8 nr 2016/22/M/NZ9/00251

Narodowe Centrum Nauki, Polska, projekt: Opus 12 nr 2016/23/B/NZ9/02677

Obliczenia wykonywano częściowo na klastrach obliczeniowych
Poznańskiego Centrum Superkomputerowo-Sieciowego (PCSS)

Streszczenie

Współczesne metody sekwencjonowania RNA służą w przeważającej liczbie przypadków do oceny poziomu ekspresji genów i ilościowego porównywania ekspresji w różnych próbach biologicznych. Dane transkryptomiczne można jednak rozpatrywać również w kontekście jakościowym, znajdując różnice, podobieństwa i wyciągając na ich podstawie wnioski.

Pierwszym celem niniejszej pracy było opracowanie metody jakościowej analizy sekwencyjnych danych transkryptomicznych w doświadczeniach wieloczynnikowych. Drugim - dołączenie metody interpretacji efektów biologicznych do tej analizy statystycznej. Wreszcie jako trzeci cel obrane zostało sprawdzenie, czy porównanie jakościowe może również poprawić lub wprowadzić nowy kontekst do analizy ilościowej.

Aby uzyskać porównania zbiorów danych w kontekście wspólnie i rozłącznie występujących sekwencji nukleotydów, wykorzystana została odległość Mash. Jest ona podstawą prowadzonej następnie analizy zmienności jakościowej transkryptomów i identyfikacji efektów statystycznie istotnych ze względu na ich zróżnicowanie. Ponadto, zaproponowano użycie ontologii GOSlim GOA w celu uzyskania bardziej szczegółowej interpretacji biologicznej wyników. Przedstawiono przykład zastosowania metody na rzeczywistych danych eksperymentalnych oraz omówiono uzyskane tym sposobem wyniki.

Rzeczywiste dane stanowiły dane transkryptomiczne z eksperymentu trójczynnikowego (czynniki: genotyp, dzień, temperatura), przeprowadzonego na jęczmieniu (Mikołajczak i in., 2022), uzyskane przez sekwencjonowanie za pomocą platformy Illumina. Dane te zostały poddane kontroli jakości i składaniu w skafoldy. Dla każdej pary zbiorów została obliczona odległość Mash, w wyniku czego otrzymano macierz odległości, która została poddana analizie PCoA, a następnie jedno- i wielowymiarowej analizie wariancji. Obliczenia przeprowadzone zostały dla surowych danych, dla skafoldów oraz wspólnie dla zbioru surowych i złożonych danych, w celu porównania wyników. Wyniki analiz wykazały największy wpływ czynnika temperatura oraz czynnika dzień na różnice jakościowe w zbiorach transkryptomów. Etap składania zauważalnie różnicował zbiory między sobą, natomiast wyniki osobno dla skafoldów i dla surowych danych były podobne.

Zbiory danych złożonych zostały podzielone na podzbiory ze względu na terminy GOSlim przypisane do poszczególnych sekwencji. Dla każdego terminu, dla którego podzbiory danych były niepuste, przeprowadzona została analiza jak ta opisana powyżej. Podobnie jak dla całych zbiorów, dla podzbiorów na ogół wykazano istotności efektów czynników: temperatury i dnia.

Dane zostały poddane różnicowej analizie ilościowej za pomocą dwóch modeli: modelu liniowego analizy wariancji oraz z wykorzystaniem modelu analizy kowariancji, gdzie zmiennymi objaśniającymi były współrzędne główne. Liczby genów wykazujących zróżnicowany poziom ekspresji uzyskane przy wykorzystaniu informacji zawartych w odległościach Mash, tzn. w modelu analizy kowariancji, były niższe niż te uzyskane przy użyciu modelu liniowego.

Najważniejsze wnioski ogólne wynikające z przedstawionej pracy to: odległość Mash jest odpowiednim narzędziem do wprowadzenia kontekstu jakościowego do analizy danych transkryptomicznych, opracowana metoda jest stosunkowo mało wymagająca w kontekście czasu i zasobów, a zastosowanie ontologii GOSlim pozwoliło na wzbogacenie analizy danych transkryptomicznych o element funkcjonalny. Wykorzystanie wyników z PCoA wykonanej na macierzy odległości Mash umożliwiło poprawienie wyników różnicowej analizy ilościowej.

Słowa kluczowe: odległość, Mash, transkryptom, GO, GOSlim, doświadczenia wieloczynnikowe, interakcja genotypu i środowiska

Abstract

In most cases, modern RNA sequencing methods are used to assess the level of gene expression and quantitatively compare expression in different biological samples. However, transcriptomic data can also be considered in a qualitative context, when we find differences, similarities and draw conclusions based on them.

The first goal of this work was to develop a method for qualitative analysis of transcriptomic sequence data in multifactorial experiments. The second goal was to add a method for interpreting biological effects to this statistical analysis. Finally, the third goal was to check whether qualitative comparison can also improve or introduce a new context to quantitative analysis.

To obtain comparisons of data sets - in the context of jointly and disjointly occurring nucleotide sequences - the Mash distance was used. It is the basis for the subsequent analysis of qualitative variability of transcriptomes and identification of statistically significant effects due to their diversity. Furthermore, the use of the GOSlim GOA ontology was proposed to obtain a more detailed biological interpretation of the results. Examples of application of the methods on real experimental data are presented and the results obtained in this way are discussed.

Real transcriptomic data from a three-factor experiment (factors: genotype, day, temperature), conducted on barley (Mikołajczak i in., 2022) were obtained by Illumina sequencing. These data were subjected to quality control and assembled into scaffolds. For each pair of sets, the Mash distance was calculated, resulting in a distance matrix that was subjected to PCoA analysis and then univariate and multivariate analysis of variance. Calculations were performed for raw data, for scaffolds, and jointly for the raw and assembled data sets, in order to compare the results. The results of the analyses showed the greatest influence of the temperature factor and the day factor on the qualitative differences in the transcriptome sets. The assembly step noticeably differentiated the sets from each other, while the results for scaffolds and raw data separately were similar.

The assembled data sets were divided into subsets based on the GOSlim terms assigned to the individual sequences. For each term for which the subsets of data were non-empty, the analysis described above was performed. As for the entire sets, the effects of the temperature factor and the day factor were significant.

The data were subjected to quantitative differential analysis using two models: a linear analysis of variance model and an analysis of covariance model, where the explanatory variables were the principal coordinates. The numbers of genes showing differential expression levels obtained using the information contained in the Mash distances (i.e., in the analysis of covariance model) were lower than those obtained using the linear model.

The most important general conclusions resulting from the presented work are: Mash distance is a suitable tool for introducing a qualitative context to the analysis of transcriptomic data, the developed method cost is relatively low in terms of time and resources, and the use of the GOSlim ontology allowed enriching the analysis of transcriptomic data with a functional element. Using the results from PCoA performed on the Mash distance matrix allowed improving the results of differential quantitative analysis.

Keywords: distance, Mash, transcriptome, GO, GOSlim, factorial experiments, genotype by environment interaction

1. Wstęp

W czasach, w których dzięki nieustannie poszerzanej gamie przyrządów pomiarowych oraz innych technicznych rozwiązań mamy coraz więcej możliwości prowadzenia obserwacji otaczającego nas świata, rośnie również zapotrzebowanie na opracowywanie metod pozwalających sprawnie analizować i systematyzować ogromne ilości danych, a także wyciągać z nich jak najwięcej przydatnych informacji. Nauki biologiczne nie są tutaj wyjątkiem.

W kontekście roślin często spotykanym modelem eksperymentu jest obserwacja wpływu kilku czynników doświadczalnych na organizmy poprzez fenotypowanie oraz zastosowanie co najmniej jednej z technik omicznych, np. genomiki, transkryptomiki, proteomiki lub metabolomiki. Badanie więcej niż jednego czynnika pozwala nie tylko zmniejszyć cenę i koszt czasowy badań, ale też wychwycić pewne zależności – interakcje między czynnikami – których bez tego szerszego spojrzenia nie dałoby się zauważyć. Eksperymenty wieloczynnikowe nie są nowym wynalazkiem. Możliwości sekwencjonowania i wykrywania cząsteczek w takich układach doświadczalnych oraz narzędzia bioinformatyczne do obróbki i analizy otrzymanych danych rozwijają się bardzo dynamicznie.

W przypadku eksperymentów wieloczynnikowych, w których badany jest transkryptom, najczęściej stosowaną praktyką po otrzymaniu danych sekwencyjnych jest: sprawdzenie ich jakości, usunięcie błędnych danych, złożenie w sekwencje konsensusowe, a następnie analiza ilościowa, pozwalająca porównać odpowiedzi badanych organizmów na kombinacje poziomów czynników i stwierdzić, czy efekty czynników były, czy też nie były istotne statystycznie w sensie zmiany poziomu ekspresji genów. Ze względu na możliwość określania liczebności występowania konkretnych transkryptów, analiza ilościowa stała się domeną eksperymentów na roślinach wykorzystujących protokół sekwencjonowania nowej generacji RNA (RNA-seq). Zdecydowanie rzadziej w tego typu badaniach pojawia się kontekst jakościowy.

Zasadną wydaje się więc chęć eksploracji i rozwinięcia metod związanych z jakościową analizą porównawczą danych transkryptomicznych otrzymanych w wyniku doświadczeń wieloczynnikowych. Do spełnienia tego celu zdecydowaliśmy się na wykorzystanie odległości Mash, która ocenia względny udział części wspólnej badanych zbiorów sekwencji.

W pracy, w rozdziale drugim, zaprezentowany został stan wiedzy na temat sekwencjonowania transkryptomów, natury zmian sekwencji nukleotydowych oraz metod badania zmienności nukleotydowej – w ramach tego ostatniego przedstawiona została idea korzystania z pojęcia odległości oraz opisana została odległość Mash. W rozdziale trzecim nakreślony został cel pracy: zaproponowanie nowego sposobu analizy danych transkryptomicznych, opartego na odległości. Wyszczególniono również wymagania, które powinna taka analiza spełniać. W rozdziale czwartym przedstawione zostały metody statystyczne i bioinformatyczne oraz narzędzia obliczeniowe wykorzystane w pracy. W rozdziale piątym znaleźć można informacje na temat danych doświadczalnych wykorzystanych w celu sprawdzenia działania opracowanej metody, a w rozdziale szóstym – wyniki uzyskane dla tych danych za pomocą opisanej metody analizy. W siódmym rozdziale zamieszczona została dyskusja poruszająca tematy, które uznane były za konieczne do omówienia lub warte poruszenia w związku z zaproponowanymi metodami i otrzymanymi wynikami, oraz wnioski końcowe.

2. Stan wiedzy

2.1. Sekwencjonowanie DNA

Historia rozwoju technologii sekwencjonowania zawiera w sobie pracę bardzo wielu ludzi oraz odkrycia, które stopniowo przybliżały nas do obecnego stanu wiedzy i umiejętności. Poniższy zarys przytacza część z tych dokonań (za rysem historycznym z: Dahm, 2005 oraz https://www.nature.com/scitable/ebooks/cntNm-8/118521938/).

W 1869 roku Johannes Friedrich Miescher po raz pierwszy wyizolował kwasy nukleinowe, początkowo nazywane nukleinami. Parępaście lat później Albrecht Kossel odkrył, że kwasy te składają się z pięciu zasad azotowych (www.nobelprize.org a), a także wyizolował trzy z nich: adeninę (1885), tyminę (1893) i cytozynę (1894). Guanina wyodrębniona została dużo wcześniej, bo w 1844 roku, przez Juliusa Bodo Ungera, z odchodów ptaków morskich, czyli guana - stąd też jej nazwa. Najpóźniej, bo w 1900 roku, odkryty został uracyl, wyizolowany przez Alberto Ascoli'ego w 1900 roku.

Kolejną ważną datą dla historii sekwencjonowania był rok 1928, kiedy to Frederick Griffith przeprowadził eksperyment na myszach, z wykorzystaniem bakterii *Streptococcus pneumoniae* (Griffith, 1928). Jego celem było opracowanie szczepionki przeciw zapaleniu płuc – uzyskał jednak inny zaskakujący wynik i pierwszą, chociaż wtedy nie do końca zrozumianą, obserwację zjawiska przyjęcia przez bakterie „cząstki transformującej". To zachęciło trzech naukowców: Oswalda Avery'ego, Colina MacLeoda i Maclyna McCarty'ego do dalszego zbadania tego zjawiska. W 1944 roku opublikowali oni wyniki z przeprowadzonego doświadczenia, w którym otrzymali argumenty przemawiające za tym, że to DNA jest ową „cząstką transformującą" (Avery i in., 1944). Pozostali jednak ostrożni przy wyciąganiu wniosków i dopiero w 1952 roku Martha Chase i Alfred Hershey eksperymentalnie potwierdzili tę tezę (Hershey i Chase, 1952).

Następną zagadką do rozwiązania był kształt DNA. W 1938 roku William Astbury i Florence Bell opublikowali w *Nature* artykuł opisujący jego strukturę jako „stos groszy" (ang. „Pile of Pennies"; Astbury i Bell, 1938). W 1952 Rosalind Franklin pracowała nad uzyskaniem danych krystalograficznych o DNA, w wyniku czego powstało między innymi słynne „Zdjęcie 51" (ang. "Photo 51"), czyli zdjęcie rentgenowskie przeprowadzonej przez nią dyfrakcji, wykonane przez studenta, Raymonda Goslinga (Franklin i Gosling, 1953). Fotografia ta została przekazana przez Maurice'a Wilkinsa Jamesowi Watsonowi i Francisowi Crickowi, którzy korzystając z zawartych w niej informacji i wyników własnych badań, w 1953 roku zaproponowali trójwymiarowy model podwójnej helisy DNA (Wattson i Crick, 1953).

Zatem w połowie XX wieku ludzkość wiedziała już, gdzie się znajduje informacja genetyczna, z czego jest zbudowana i jaką ma strukturę przestrzenną. Jak ją jednak rozszyfrować? Potrzeba było kolejnych 12 lat obserwacji i udoskonaleń technik laboratoryjnych, by w roku 1965 Robert Holley odczytał kolejność nukleotydów w wyizolowanej z drożdży cząsteczce transportującego RNA (tRNA; Holley i in., 1965) i następnych 12 lat, aby w 1977 roku, w dwóch pracowniach – Fredericka Sangera i Waltera Gilberta – niezależnie opracowano dwa sposoby sekwencjonowania DNA. Sanger opracował metodę enzymatyczną (Sanger i in., 1977), natomiast Maxam i Gilbert – opartą na degradacji chemicznej (Maxam i Gilbert, 1977).

Najwcześniejsze metody sekwencjonowania nazywamy sekwencjonowaniem pierwszej generacji. Pozwalają one na dokładny odczyt jednej zadanej sekwencji. Należąca do tego zbioru technik metoda

Sangera – po modyfikacjach związanych z odkryciem przez Karry'ego Mullisa (www.nobelprize.org b) reakcji łańcuchowej polimerazy (PCR) – używana jest do dziś. Zaletami tego sekwencjonowania jest wysoka dokładność (poza początkiem sekwencji o długości 15 do 40 nukleotydów, czyli poza miejscem przyłączania się startera) oraz stosunkowo długie odczyty – od 300 do 1000 nukleotydów. Ma jednak również pewne ograniczenia. W jednej reakcji poznajemy jedną sekwencję, co w niektórych przypadkach w zupełności nam wystarczy, jednak kiedy do zbadania mamy więcej materiału genetycznego, zadanie staje się pracochłonne i drogie. W odpowiedzi na rosnące potrzeby powstały metody zwane sekwencjonowaniem drugiej generacji - ich początki związane są z opracowaną na przełomie wieków XX i XXI technologią firmy Solexa, przejętą w roku 2007 roku przez firmę Illumina.

Sekwencjonowanie drugiej generacji, nazywane też sekwencjonowaniem następnej generacji (ang. Next Generation Sequencing, NGS), pozwala na równoległe odczytywanie sekwencji wielu fragmentów DNA. Jest stosunkowo tanie i wydajne. Długość odczytów zazwyczaj zawiera się w przedziale od 50 do 300 nukleotydów, co wynika z faktu, że w trakcie przygotowania do sekwencjonowania materiał zostaje poddany fragmentacji. Sekwencje odczytywane są od końców 5' i 3', w wyniku czego otrzymuje się tzw. lewe i prawe odczyty. Przy pomocy odpowiednich narzędzi bioinformatycznych przeprowadza się składanie odczytów w dłuższe sekwencje (ang. assembly). Przykładem narzędzia wykorzystywanego w tym procesie jest SOAPdenovo2, który bardziej szczegółowo omówimy rozdziale 4.1.2, jednak ogólną ideą jest znajdowanie wśród odczytanych ciągów nukleotydów wspólnych, nakładających się fragmentów i składanie z nich dłuższych, konsensusowych ciągów. Niewielka długość i potrzeba składania odczytów generuje pewien problem – sekwencjonowanie II generacji przestaje być efektywnym rozwiązaniem, kiedy stajemy przed wyzwaniem poznawania całych nowych genomów czy też transkryptomów. Metody składania nie radzą sobie dobrze z regionami powtarzających się sekwencji. Potrzebne więc były nowe rozwiązania - jednym z nich jest sekwencjonowanie III generacji.

Sekwencjonowanie III generacji pozwala na odczytywanie długich sekwencji nukleotydowych. Przykładami technologii z tej grupy dostępnych obecnie na rynku są metody: SMRT (Single Molecule Real Time) oferowana przez Pacific Biosciences (PacBio) oraz Oxfordzki MinION (Oxford Nanopore Technologies).

2.2. Natura i badanie zmian sekwencji nukleotydowych RNA

Na samym początku warto podkreślić, że praca ta tworzona była z myślą o roślinach. Większość rozważań prawdopodobnie będzie również miała zastosowanie do innych organizmów, jednak w celu uniknięcia nieporozumień, już teraz sygnalizujemy ten fakt.

Powstanie metod wielkoskalowego sekwencjonowania DNA umożliwiło naukowcom efektywne podejście do poznania jakościowej zmienności sekwencji charakterystycznych dla różnych organizmów. W tym obszarze badań stosowane jest najczęściej sekwencjonowanie genomów. Różnice pomiędzy organizmami są interpretowane w odniesieniu do ich filogenezy, czyli wspólnych przodków i mechanizmów powstawania gatunków. Możliwe jest też badanie różnic powstałych w krótszych okresach czasowych pod wpływem czynników środowiskowych i mających postać mutacji punktowych.

Drugim obszarem badawczym, w którym analiza zmienności sekwencji ma szerokie zastosowanie, jest porównywanie zbiorów sekwencji RNA powstałych z genomowego DNA na drodze transkrypcji. Cząsteczki RNA pełnią różne funkcje, jednak najliczniejszą grupę stanowią takie, które zawierają informacje umożliwiające syntezę białek z aminokwasów w wyniku procesu zwanego translacją (ang. translation). Taki RNA nazywa się informacyjnym RNA, w skrócie mRNA, od angielskiej nazwy:

messenger RNA (Alberts i in., 2019). Cząsteczki RNA są znacznie bardziej zmienne w sensie jakościowym niż cząsteczki genomowego DNA. Ponadto podlegają one zmienności ilościowej, która świadczy o zmiennej ekspresji genów. Mówiąc o analizach prowadzonych na zsekwencjonowanych cząsteczkach RNA, mówimy o zsekwencjonowanym transkryptomie. Transkryptom to zestaw wszystkich RNA zawartych w danej komórce, grupie komórek lub całym organizmie w jednym punkcie czasowym. Te metadane – punkt czasowy i źródło materiału – są niezbędne do prowadzenia analiz porównawczych i wyciągania właściwych wniosków.

W sensie technicznym sekwencjonowanie transkryptomu sprowadza się do sekwencjonowania DNA komplementarnego do interesującego nas RNA (ang. copy/complementary DNA, cDNA). W celu otrzymania cDNA, izoluje się mRNA (korzystając z faktu, że do mRNA w trakcie obróbki posttranskrypcyjnej dołączane są na końcu łańcuchy składające się z nukleotydów adeninowych, czyli tak zwane ogony poliA), sprawdza pod kątem stanu degradacji i traktuje enzymem zwanym odwrotną transkryptazą. Otrzymane tym sposobem cząsteczki są trwalsze, co pozwala na odczytanie ich sekwencji (Deonier i in., 2005).

Jakościowe i ilościowe porównywanie transkryptomów dostarcza ważnych informacji na temat badanych organizmów - może na przykład wskazać konkretne szlaki metaboliczne, mechanizmy odpowiedzialne za odporność czy też podpowiedzieć, co wpływa na różnice fenotypowe pomiędzy badanymi organizmami. Żeby dobrze interpretować obserwowane różnice w transkryptomach, musimy wiedzieć, co może odpowiadać za ich zmienność.

Pierwsze czynniki to genom, czyli zbiór wszystkich genów właściwych dla danego organizmu, oraz genotyp, czyli sekwencje genów właściwe dla danego osobnika (warianty genów). U organizmów posiadających różne genomy (na przykład dąb i dziurawiec) lub genotypy (na przykład dwie odmiany pszenicy) możemy spodziewać się różnic w transkryptomach, zarówno jakościowych jak i ilościowych (chociaż mogą też zaistnieć warunki, kiedy transkryptomy nie będą się znacząco różniły). Różnice te wynikają z przystosowania organizmów do życia w danych warunkach na drodze filogenezy lub działania człowieka (inżynieria genetyczna, hodowla i selekcja).

Zakłada się, że genom organizmu jest taki sam we wszystkich jego tkankach, jednak występują wyjątki od tej reguły. Przykładami są tzw. mozaikowość genetyczna oraz chimeryzm. Mozaikowość to zjawisko polegające na występowaniu różnic genetycznych w organizmie powstałym z dwóch komórek rozrodczych. Rozróżnia się mozaikowość prawdziwą i funkcjonalną. Przyczyną pierwszej mogą być punktowe mutacje, rekombinacje mitotyczne, utrata bądź „pozyskanie" chromosomu. Druga występuje na skutek zmiany pozycji genu (position-effect variegation, PEV), interferencji RNA, inaktywacji chromosomu X lub inaktywacji poszczególnych genów (piebaldyzm, heterochromia) (Szabad, 2021). Chimeryzm z kolei dotyczy organizmów, które powstały z więcej niż dwóch komórek rozrodczych bądź – w przypadku roślin – w wyniku zaburzenia mitozy w stożku wzrostu lub poprzez szczepienie roślin (Frank i Chitwood, 2016). Przyczyną powstania chimer może być opóźnione bądź podwójne zapłodnienie, połączenie się embrionów, iniekcja embrionów lub komórek w blastocyty czy też wymiana międzykomórkowa (Szabad, 2021).

Transkryptom może być także inny w zależności od tego, z jakiego organu (na przykład: korzeń, liść) czy też z jakiej tkanki (na przykład: okrywająca, przewodząca lub miękiszowa) będzie pobrany. W zależności od wieku organu bądź tkanki, czyli stadium rozwoju, zestaw RNA również może się różnić. Inne procesy zachodzą w młodych liściach rośliny, inne w rozwiniętych, a jeszcze inne w starych czy obumierających (Lu i in., 2020). Te różnice rozlicznych własności oraz funkcji nie mogą być wyjaśnione jedynie przez różnice genomu – tłumaczą je dopiero procesy regulacji transkrypcji i procesy

potranskrypcyjne. To dzięki nim komórki organizmów mogą odpowiadać w określony sposób na czynniki zewnętrzne i wewnętrzne (Casamassimi i Ciccodicola, 2019).

Różnice sekwencji RNA pomiędzy próbami biologicznymi powstające w czasie transkrypcji mogą wynikać z różnic w strukturze chromosomów lub chromatyny powodowanych przez działanie białek regulujących chromatynę lub zjawiska epigenetyczne (takie jak: metylacja, acetylacja itd.), jak również z różnic w oddziaływaniach pomiędzy białkami i DNA (na przykład: działanie czynników transkrypcyjnych, enhancerów).

Po transkrypcji z otrzymanego pre-mRNA (cząsteczki przepisanej z DNA, zawierającej zarówno kodujące eksony, jak i niekodujące introny) uzyskiwany jest dojrzały RNA. Procesy prowadzące do tego przekształcenia – zwane procesami potranskrypcyjnymi - również mogą powodować różnice w transkryptomie, wynikające z większej lub mniejszej trwałości RNA, dostępności cząsteczki dla czynników inicjujących translację albo po prostu z różnicy w uzyskiwanych sekwencjach. Do tej grupy procesów zalicza się między innymi: składanie (ang. splicing, polegający na wycięciu z preRNA niekodujących fragmentów i połączeniu części kodujących), edytowanie (ang. editing, zjawisko polegające na zmianie sekwencji nukleotydów na inną sekwencję w wyniku insercji, delecji lub deaminacji nukleozydów), modyfikacje RNA (ang. RNA modifications, polegające na dodawaniu cząsteczek chemicznych bezpośrednio do nukleotydów), poliadenylację oraz dołączanie czapeczki (czyli odpowiednio: dołączenie ogona poliA - łańcucha składającego się z 100-250 nukleotydów adeninowych - do końca 3’ cząsteczki pre-mRNA oraz dołączenie zmodyfikowanej guanozyny do końca 5’ pre-mRNA; elementy te decydują o odporności na degradację cząsteczek mRNA) (Prall i in., 2019).

Ostatnią grupą czynników, jaką należy tu wymienić, są czynniki zewnętrzne – między innymi: pierwiastki i związki chemiczne występujące w ziemi, wodzie lub w powietrzu, różnego rodzaju niedobory, warunki pogodowe, działanie patogenów czy też uszkodzenia mechaniczne organizmu. Ogólnie, jakościowe i ilościowe zmiany transkryptomu mogą następować pod wpływem działania czynników środowiskowych - zarówno biotycznych, jak i abiotycznych, które mogą powodować mutacje genomowe lub wpływać na opisane wcześniej procesy transkrypcyjne i potranskrypcyjne. Rozróżnienie tych dwóch grup czynników zewnętrznych znajduje swoje przełożenie w obserwowanych zmianach w transkryptomie – na przykład w *Arabidopsis thaliana* wiele genów reaguje tylko na jeden z dwóch rodzajów stresów - stresy biotyczne albo abiotyczne (Biniaz i in., 2022).

Zauważmy, że badanie zmienności sekwencji genomowego DNA pomiędzy różnymi organizmami koncentruje się zwykle na aspektach jakościowych, natomiast badanie zmian sekwencji RNA wywołanych przez mechanizmy regulacyjne lub czynniki zewnętrzne – na aspektach ilościowych. Najczęściej bowiem o badaniu zmienności genomowego DNA mówimy w kontekście: analiz filogenetycznych, testów genetycznych (oznaczanie markerów np. w genealogii, kryminalistyce, archeologii) czy testów i analiz pod kątem identyfikacji genów odporności lub wykrywania przyczyn i przypadków chorób genetycznych. Są to analizy opierające się najczęściej na aspektach jakościowych; wyjątkiem może być np. sprawdzanie liczby kopii danego genu. Natomiast badanie zmienności RNA polega najczęściej na analizie poziomu ekspresji genów lub ich izoform. Stosowane są do tego metody oparte na sekwencjonowaniu lub na reakcji łańcuchowej polimerazy, które dostarczają danych ilościowych. Analizy jakościowe RNA są wykorzystywane przy odkrywaniu genomów i transkryptomów dla gatunków, dla których jeszcze nie są one znane, poszukiwaniu miejsc splicingu, badaniu działania czynników potranskrypcyjnych, genów fuzyjnych (genów powstałych przez połączenie dwóch niezależnych genów na drodze translokacji, delecji lub inwersji chromosomowej) lub zmian pojedynczych nukleotydów bądź innych wariantów sekwencji (Yalamanchili i in., 2017;

Conesa i in., 2016). Czyni się to poprzez bezpośrednie porównanie sekwencji lub współliniowe dopasowanie badanych sekwencji. Rzadziej wykonuje się analizy jakościowe na większą skalę, na większej liczbie organizmów lub prób biologicznych, dla opisania zjawiska plastyczności RNA zapewniającej funkcjonowanie organizmów. Jak wynika z przedstawionych wyżej argumentów, zmiany takie mogą być wywołane przez szereg czynników. Poszczególne czynniki – genetyczne i środowiskowe – mogą wykazywać efekty niezależne lub też wchodzić ze sobą w interakcje. Postępy metod sekwencjonowania dają możliwość przeprowadzania doświadczeń o strukturze wieloczynnikowej, uwzględniającej wiele możliwych przyczyn zmian sekwencji RNA. Istnieje potrzeba rozwoju metod obliczeniowych i statystycznych właściwych dla tego typu badań.

2.3. Metody badania zmienności transkryptomu

Zmienność ilościową transkryptomu bada się sekwencjonując RNA wyizolowany z materiału biologicznego reprezentującego warianty uwzględnione w doświadczeniu oraz oceniając udział ilościowy poszczególnych transkryptów, różniących się sekwencjami, w całkowitym RNA. Inaczej mówiąc, poszczególne próby biologiczne opisuje się cechami o charakterze ilościowym, zawierającymi informacje o liczbie poszczególnych transkryptów. Podejście takie w naturalny sposób dostarcza obserwacji ciągłych zmiennych losowych, które następnie są poddawane analizie statystycznej metodami właściwymi dla tego rodzaju zmiennych. Przykładem jest analiza różnicowa ekspresji genów.

Analiza różnicowa pozwala na stwierdzenie, czy liczby pewnych transkryptów, czyli poziomy ekspresji pewnych genów, różnią się istotnie w kontekście porównywania prób uzyskanych w poszczególnych warunkach eksperymentalnych. Po otrzymaniu odczytów z sekwencjonowania transkryptomu (zbiorów danych z eksperymentu wieloczynnikowego, każdy zbiór opisany odpowiednimi metadanymi), kolejnymi krokami prowadzącymi do tej analizy są: kontrola jakości danych, złożenie odczytów w sekwencje konsensusowe, mapowanie w genomie referencyjnym (lub w innych sekwencjach konsensusowych będących referencją) i zliczenie występowania każdego transkryptu w zbiorze danych. Otrzymane tą drogą pliki ze zliczeniami wraz z odpowiednio przygotowanym opisem (tabela z nazwami zbiorów danych oraz przypisanymi do nich opisami traktowań) stanowią dane wejściowe dla analizy różnicowej.

Jednymi z najczęściej stosowanych narzędzi w analizie różnicowej są:

- w analizie i kontroli jakości danych - FastQC (Andrews, 2010) i AdapterRemoval (Schubert i in., 2016); inne popularne narzędzia podobne do FastQC to np. HTSeq (Anders i in., 2015), NGS QC Toolkit (Patel i Jain, 2012) czy Trim Galore (https://github.com/FelixKrueger/TrimGalore), natomiast alternatywy dla AdapterRemoval to np. Trimmomatic (Bolger i in., 2014), fastp (Chen i in., 2018) lub FASTX Toolkit (https://github.com/agordon/fastx_toolkit);
- przy składaniu odczytów – SOAPdenovo2 (Luo i in., 2012); innym narzędziem jest np. Cufflinks (https://cole-trapnell-lab.github.io/cufflinks/papers/);
- do mapowania odczytów w genomie referencyjnym - TopHat (Trapnell i in., 2009);
- do mapowania w sekwencjach konsensusowych i zliczania odczytów - Salmon (Patro i in., 2017) lub Kallisto (https://github.com/pachterlab/kallisto); innym przykładem jest Cufflinks;
- do analizy różnicowej np. funkcje DESeq2 lub edgeR w pakiecie R (Love i in., 2014).

W przypadku badania zmienności jakościowej sytuacja wygląda inaczej. Sekwencje RNA można rozumieć jako ciągi elementów ze zbioru {A, C, G, T, N} (dla uproszczenia, stosujemy symbol T dla

uracylu). Niemożliwe jest bezpośrednie określenie zmiennych ilościowych, które opisywałyby sekwencje lub całe zbiory sekwencji i pozwalałyby na ich porównywanie, przy jednoczesnym uniknięciu utraty istotnych informacji. Pojęciem, które w tej sytuacji będzie dla nas pomocne, jest „odległość".

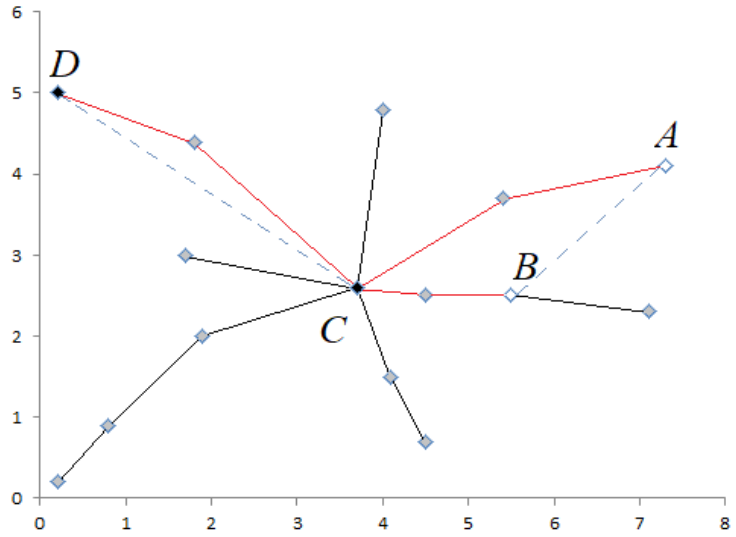
2.3.1. Odległości i ich zastosowania

Odległość to jedno z podstawowych narzędzi pomagających ludziom w badaniu otaczającego nas świata. Żeby wyznaczyć odległość między obiektami x i y należącymi do pewnego zbioru P, należy najpierw określić metrykę d, czyli funkcję przyporządkowującą każdej parze elementów z P liczbę rzeczywistą nieujemną, która spełnia warunki:

1) $d(x,x) = 0$ dla każdego elementu x z P;
2) $d(x,y) = d(y,x)$ dla każdej pary elementów x i y należących do zbioru P;
3) $d(x,y) + d(y,z) \geq d(x,z)$ dla każdej trójki elementów x, y i z należących do P.

Wówczas odległość między elementami x i y ze zbioru P to wartość $d(x, y)$ metryki d dla tych obiektów, a obiekt (P, d) nazywamy przestrzenią metryczną.

Patrząc na powyższą definicję nietrudno dojść do wniosku, że istnieje nieskończenie wiele sposobów na określenie metryki dla zadanego zbioru (na przykład mając zdefiniowaną już metrykę $d(x, y)$ można stworzyć $d^*(x, y) = a \cdot d(x, y)$, gdzie a to dowolna dodatnia liczba rzeczywista, natomiast $\cdot$ to mnożenie przez skalar) – jednak nie wszystkie funkcje określające odległości będą dla nas przydatne, tak samo jak nie wszystkie będą w ten sam sposób obrazować układ tych samych obiektów względem siebie. Przypatrzmy się przykładowi przedstawionemu na Ryc. 2.1.



Ryc. 2.1. Porównanie dwóch metryk zastosowanych na jednym zbiorze obiektów. W zależności od tego czy zastosuje się metrykę euklidesową (błękitne linie przerywane) czy modyfikowaną metra paryskiego (czerwone linie ciągłe), inna jest odpowiedź na pytanie: która para punktów – A i B czy C i D – jest bliżej siebie?

Na płaszczyźnie $\mathbb{R}^2$ zaznaczone są wszystkie punkty zbioru P , będącego skończonym podzbiorem elementów z $\mathbb{R}^2$, oraz wszystkie odcinki ze zbioru O, będącego podzbiorem zbioru odcinków łączących punkty ze zbioru P. Zauważmy, że niektóre pary punktów ze zbioru P są połączone odcinkami ze zbioru

O, a inne nie. Elementami zbioru *P* są między innymi punkty: *A, B, C* i *D*. Jeśli zadamy sobie pytanie, która para punktów jest bliżej siebie: *A* i *B* czy *C* i *D*, odpowiedź będzie brzmiała: to zależy od metryki, jaką przyjmiemy.

Weźmy pod uwagę dwie metryki:

- metryka euklidesowa – oznaczmy ją $e(x,y)$
- zmodyfikowana metryka metra paryskiego – oznaczmy ją $p(x,y)$

Pierwsza jest dobrze nam znaną metryką, która w przestrzeni R^2 dla dowolnych dwóch punktów $x = (x_1,x_2)$ i $y = (y_1,y_2)$ wyrażona jest wzorem:

$$e(x,y) = \sqrt[2]{(y_1 - x_1)^2 + (y_2 - x_2)^2},$$

i tak samo będzie ona działała na podzbiorze $\mathbb{R}^2$stanowiącym zbiór punktów *P*.

Tę drugą również zdefiniujemy dla zbioru P: niech odległością między dowolnymi dwoma punktami dla powyższego układu punktów będzie wyrażona przez minimalną liczbę odcinków ze zbioru *O*, które wraz z pewnymi wierzchołkami z *P* tworzą łamaną łączącą oba punkty, czyli:

$$p(x,y) = \min\{|t|, t \in A\},$$

gdzie A to zbiór łamanych, których wszystkie odcinki są zawarte w O, a w zbiorze wierzchołków zawarte są punkty x i y, natomiast $|t|$ to liczba odcinków zawartych w łamanej t. Szybko można sprawdzić, że definicja ta spełnia warunki metryki.

Mając powyższe dwie definicje możemy wrócić do pytania: które dwa punkty są mniej oddalone od siebie: *A* i *B* czy *C* i *D*? Zauważmy, że jeśli przyjmiemy dla zbioru *P* metrykę euklidesową, to odcinek łączący A i B jest krótszy niż odcinek łączący C i D, to znaczy:

$$e(A,B) < e(C,D),$$

czyli odległość między parą *A* i *B* jest mniejsza niż między *C* i *D*. Jednak w przypadku zmodyfikowanej metryki paryskiej mamy:

$$2 = p(C,D) < p(A,B) = 4,$$

czyli to para *C* i *D* jest bliżej niż *A* i *B*.

Ten przykład pokazuje, że pojęcie „daleko" i „blisko" są względne – zależą od metryki, jaką się przyjmie. Od czego zależy wybór metryki? Od osoby badającej odległości między danymi obiektami – to ona musi zadać sobie pytanie, jaka (w jaki sposób liczona) odległość dobrze zobrazuje rzeczywisty układ obiektów względem siebie i pozwoli na porównanie ich konkretnych cech.

2.3.2. Odległości między sekwencjami

Odległości znajdują swoje zastosowania w wielu dziedzinach nauki, również w badaniu różnic między sekwencjami. Zdefiniowanie odpowiedniej metryki nie zawsze jest jednak oczywiste i nie ma jednoznacznej odpowiedzi na pytanie, którą najlepiej zastosować. Zależy to między innymi od rodzaju danych, którymi dysponujemy, pytania, na które poszukujemy odpowiedzi oraz dostępnego czasu i mocy obliczeniowej.

Intuicyjnie najprostszą metryką, zdefiniowaną na zbiorze sekwencji tej samej długości, jest odległość Hamminga, wprowadzona przez Richarda Hamminga w 1950 roku. Dla każdych dwóch sekwencji nukleotydów zlicza ona liczbę miejsc, w których odpowiadające sobie nukleotydy różnią się między sobą.

AAGAATTA CAGAATTA	AAGAATTA AAGTGCCG	AAGAATTA ACAGAATT
odległość = 1	odległość = 5	odległość = 5

Ryc. 2.2. Odległość Hamminga między dwiema pierwszymi sekwencjami wynosi 1. Odległość zarówno dla drugiej jak i trzeciej pary sekwencji jest równa 5, mimo że trzecia para różni się tylko dwoma nukleotydami (wstawionym C i usuniętym A).

Zaletami tej metryki są prostota i szybkość wykonywania obliczeń, jednak nie pozwala ona na porównywanie sekwencji różnej długości. Nie jest ona również właściwa do opisywania takich zjawisk jak duplikacja, insercja czy delecja, które w przypadku tej metryki mogą diametralnie zmienić wynik (przykład: Ryc. 2.2) - a nie zawsze tego oczekujemy porównując dwie sekwencje.

Przykładem metryki pozwalającej liczyć odległości na zbiorze sekwencji różnych długości jest metryka Damareau-Levenshteina, zdefiniowana na zbiorze skończonych ciągów znaków – napisów - za pomocą pojęcia działań prostych. Jeśli działaniem prostym na napisie będzie jedna z operacji:

1) wstawienie w dowolnym miejscu nowego znaku do napisu;
2) usunięcie dowolnego znaku z napisu;
3) zamiana znaku w napisie na dowolny inny znak;
4) zamiana miejscami dwóch znaków obok siebie;

to wówczas odległością Damareau-Levenshteina $L(A, B)$ między napisami A i B będzie minimalna liczba działań prostych potrzebnych, aby jedną sekwencję przekształcić w drugą. Zaletami tej metryki jest możliwość porównywania sekwencji różnej długości oraz fakt, że algorytm ma na uwadze zjawiska powszechnie występujące w sekwencjach biologicznych. Wadą jest złożoność algorytmu, a co za tym idzie – czas potrzebny do wykonania obliczeń.

2.3.3. Odległość Mash

Zanim omówimy następną koncepcję – odległość Mash – musimy przypomnieć sobie następujące pojęcia: k-mer i indeks Jaccarda.

***k* -mer**

Mając zadany ciąg znaków S, k-merem nazywamy każdy podciąg składający się z kolejnych elementów ciągu S, którego długość wynosi k. Na przykład dla ciągu S = „ACTATGAT” 3-merami będą między innymi ACT i GAT, a 5-merami będą ACTAT czy też CTATG.

Indeks Jaccarda

Indeks Jaccarda to wartość wyznaczana dla par zbiorów, zależna od stopnia ich podobieństwa w sensie posiadania części wspólnej. Jeśli zadane są dwa zbiory elementów: A i B, indeksem Jaccarda nazwiemy iloraz:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

gdzie $|A \cap B|$ to liczba elementów wspólnych dla zbiorów A i B, a $|A \cup B|$ to liczba wszystkich elementów zawartych w zbiorze A lub zbiorze B.

Dla przykładu – niech A będzie zbiorem liczb: {1,2,3}, natomiast B niech będzie równe {2,3,4,5,6}. Część wspólna tych zbiorów to zbiór {2,3}, natomiast wszystkie elementy należące do co najmniej jednego z nich to {1,2,3,4,5,6}. Stąd indeks Jaccarda dla zbiorów A i B wynosi:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{1,2,3\} \cap \{2,3,4,5,6\}|}{|\{1,2,3\} \cup \{2,3,4,5,6\}|} = \frac{|\{2,3\}|}{|\{1,2,3,4,5,6\}|} = \frac{2}{6} = \frac{1}{3}.$$

Lemat 1.

Dla dowolnych dwóch zbiorów wartość indeksu Jaccarda należy do przedziału [0,1].

Dowód:

Załóżmy, że lemat nie zachodzi. Znaczyłoby to, że istnieje para zbiorów A i B, dla których wartość indeksu Jaccarda jest ujemna lub większa od 1. Ujemność indeksu J(A, B) możemy od razu wykluczyć, ponieważ $|A \cap B|$ oraz $|A \cup B|$ jako liczebności zbiorów nie mogą być ujemne. Pozostaje zastanowić się, kiedy iloraz powyższych byłby większy od 1: stałoby się tak w momencie, w którym liczba elementów wspólnych zbiorów A i B byłaby większa od sumy wszystkich elementów tych dwóch zbiorów - jest to jednak niemożliwe, ponieważ każdy element należący do $A \cap B$ musi należeć również do $A \cup B$. Zatem lemat musi zachodzić.

□

Odległość Mash została zdefiniowana przez Ondova i in. (Ondov i in., 2016), którzy zaproponowali algorytm o tej samej nazwie, liczący tę odległość. W ninejszej pracy nazwa Mash będzie poprzedzana słowem „odległość" lub „algorytm", żeby jednoznacznie je rozróżniać. Omówimy teraz ideę stojącą za odległością Mash oraz algorytmem Mash, liczącym jej wartości.

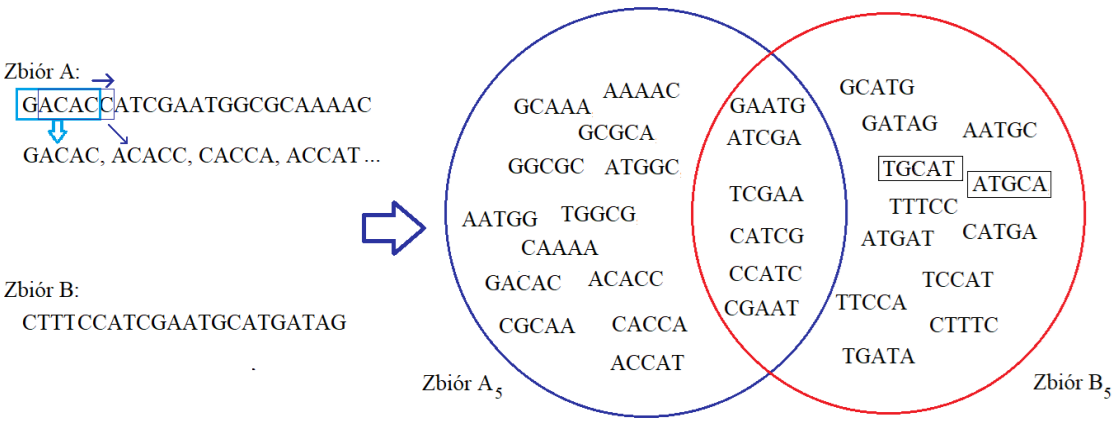
Załóżmy, że mamy dane dwa zbiory sekwencji: zbiór A i zbiór B. Oba zawierają jeden lub więcej ciągów znaków ze zbioru {A,C,G,T}, o niekoniecznie równych długościach. Ponieważ indeks Jaccarda określa podobieństwo między zbiorami, chcielibyśmy go użyć przy wyznaczaniu odległości między A i B. Żeby jednak móc go zastosować na zadanych zbiorach, najpierw ich elementy muszą być zdefiniowane w taki sposób, aby możliwe było porównywanie ich między sobą i stwierdzenie, czy są różne, czy identyczne.

Przykładem takich elementów są k-mery: dla ustalonego k dwa k-mery: $x_1 \ldots x_k$ i $y_1 \ldots y_k$ są sobie równe wtedy i tylko wtedy, gdy: $x_1 = y_1$, $x_2 = y_2, \ldots$, $x_k = y_k$. Aby przekształcić zbiory A i B w zbiory k-merów w nich występujących można wykorzystać metodę tzw. „przesuwanego okna" (ang. sliding window), polegającą na czytaniu po kolei fragmentów sekwencji długości k: od pierwszego do k-tego znaku, następnie od drugiego do (k+1)-go i tak dalej.

Tworząc reprezentację dwóch zbiorów sekwencji w postaci odpowiadających im zbiorów k-merów będzie można obliczyć przybliżony indeks Jaccarda, a następnie odległość Mash ze wzoru:

$$(1)\quad D = -\frac{1}{k} ln \frac{2j}{1+j},$$

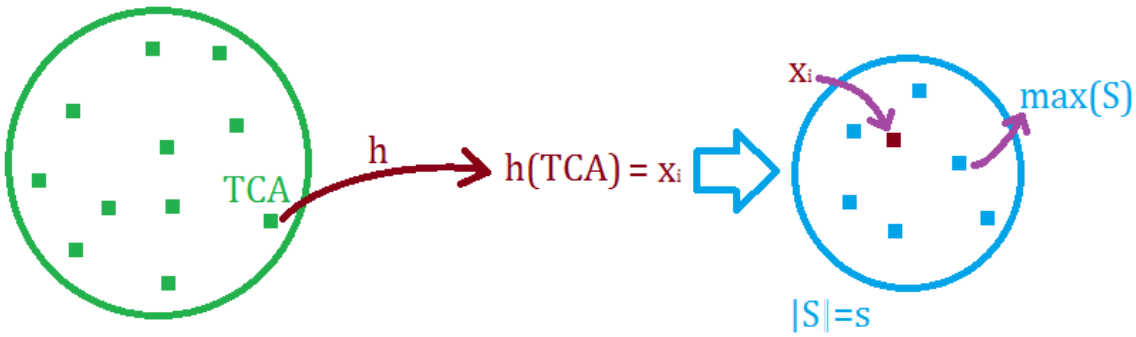
gdzie k to zastosowana długość k-merów, a j to wyznaczona przybliżona wartość indeksu Jaccarda.



Ryc. 2.3. Przykład wyznaczania 5-merów za pomocą „przesuwanego okna".

Konstrukcję zbioru *k*-merów za pomocą „przesuwanego okna" ilustruje Ryc. 2.3 na przykładzie $k = 5$. Wynikiem tego procesu są zbiory A_5 i B_5, dla których można wyznaczyć część wspólną i sumę zbiorów. Zauważmy, że zbiory *A* i *B* mogą zawierać *k*-mery powtarzające się więcej niż jeden raz oraz pary odwrotnie komplementarnych *k*-merów (np. dwa *k*-mery zaznaczone czarną ramką - ATGCA i TGCAT); takie *k*-mery powinny być uwzględniane tylko jeden raz.

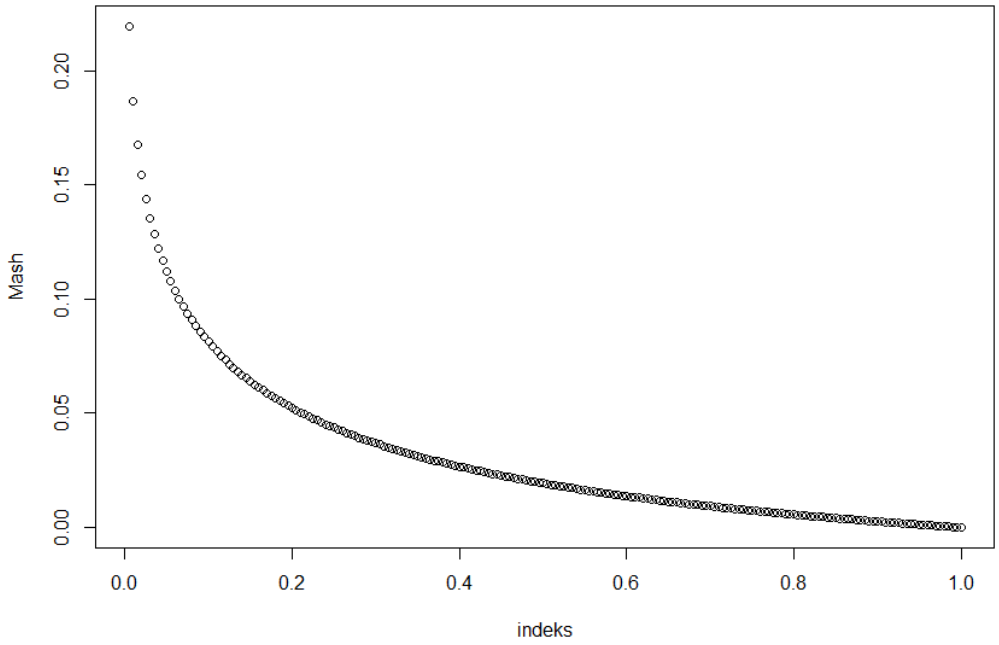
Algorytm Mash korzysta z tej metody, jednak z pewną modyfikacją pozwalającą na skrócenie czasu potrzebnego do wykonania obliczeń. W trakcie czytania kolejnych *k*-merów występujących w zbiorach *A* i *B*, algorytm stosuje funkcję haszującą *h*, która pozwala skonstruować zbiory $h(A_k)$ oraz $h(B_k)$ reprezentujące zbiory *A* i *B* – tak zwane szkice (ang. sketch). Funkcja haszująca jest narzędziem pozwalającym algorytmowi Mash działać szybko nawet na ogromnych zbiorach danych. Każdemu odczytanemu *k*-merowi przypisuje ona tak zwany hash - losową wartość 32- lub 64-bitową (zależnie od wielkości *k*) – a następnie sprawdza dla zadanego wcześniej parametru *s*, czy ten hash należy do *s* najmniejszych wartości przypisanych do *k*-merów. W ten sposób dla każdego ze zbiorów sekwencji algorytm Mash wybiera losowo *s* hashy reprezentujących *k*-mery w nim występujące (Ryc. 2.4). Na zbiorach haszy liczony jest indeks Jaccarda, który stanowi przybliżenie indeksu Jaccarda dla całych zbiorów k-merów. Następnie obliczana jest odległość Mash według wzoru (1).



Ryc. 2.4. Działanie funkcji haszującej *h*. 3-merowi TCA zostaje przyporządkowana wartość x_i, następnie sprawdzany jest warunek, czy jest ona większa od największej wartości w zbiorze *S*, zawierającym s najmniejszych haszy. Jeśli tak jest, x_i trafia do zbioru *S*, natomiast największa wartość w zbiorze, *max*(*S*), zostaje z niego usunięta.

Ponieważ indeks Jaccarda jest liczony na losowo wybranych zbiorach k-merów reprezentujących porównywane zbiory sekwencji, mówimy o przybliżonym indeksie Jaccarda dla tych zbiorów. Przy obliczeniach wyznaczane są dodatkowe parametry określające precyzję wyliczanej oceny odległości. Podawana jest liczba wspólnych haszy zawartych w szkicach oraz wartość prawdopodobieństwa przypadkowego otrzymania takiej lub większej liczby wspólnych haszy. Mała wartość tego prawdopodobieństwa oznacza wyższą wiarygodność wyniku.

Przy zadanych parametrach *s* i *k*, można wyznaczyć maksymalną wartość odległości dla pary zbiorów posiadających część wspólną. Na przykład, przy $s = 1000$ i $k = 21$, wartość ta wynosi w przybliżeniu 0,329. Zależność pomiędzy wartościami indeksu Jaccarda a wartościami odległości Mash obrazuje Ryc. 2.5.



Ryc. 2.5. Wartości odległości Mash między zbiorami danych w zależności od indeksu Jaccarda dla tych zbiorów. Za wartość k przyjęto 21. Zauważmy, że dla indeksu Jaccarda równego 0 mamy asymptotę i wartość funkcji dąży do nieskończoności.

Wykażemy teraz, że odległość Mash spełnia warunki definiujące pojęcie odległości.

Nieujemność:

Dowód tej własności przeprowadzimy metodą nie wprost. Załóżmy, że dla pewnych zbiorów sekwencji *A* i *B* zachodzi warunek:

$$D(A,B) = -\frac{1}{k} ln \frac{2J(A,B)}{1+J(A,B)} < 0$$

Żeby tak było, ułamek $\frac{2J(A,B)}{1+J(A,B)}$ musiałby mieć wartość większą od 1. Mając na względzie fakt, że indeks Jaccarda jest nieujemny, otrzymujemy nierówność:

$$2J(A,B) > 1 + J(A,B)$$

która prowadzi do nierówności $J(A,B) > 1$, która jest sprzeczna z Lematem 1. Zatem prawdą jest, że dla dowolnych dwóch zbiorów *A* i *B* wartość $D(A,B)$ jest nieujemna.

□

Identyczność nierozróżnialnych:

Ustalmy pewne dodatnie *k*. Sprawdźmy, czy zachodzi warunek:

$$D(A,B) = 0 \leftrightarrow A = B.$$

Zaczniemy od implikacji: $A = B \rightarrow D(A,B) = 0$. Ponieważ $A = B$, mamy:

$$J(A,B) = J(A,A) = \frac{|A \cap A|}{|A \cup A|} = \frac{|A|}{|A|} = 1.$$

Zatem dla dowolnego ustalonego *k*:

$$-\frac{1}{k} ln \frac{2J}{1+J} = -\frac{1}{k} ln 1 = 0.$$

Przejdźmy do implikacji $D(A,B) = 0 \rightarrow A = B$. Załóżmy, że $D(A,B) = 0$. Oznacza to, że $-\frac{1}{k} ln \frac{2J}{1+J} = 0$, co może zachodzić tylko w przypadku, gdy $\frac{2J}{1+J} = 1$, czyli gdy $J = 1$. Taki warunek zachodzi, gdy zbiory *A* i *B* są sobie równe, co kończy dowód na identyczność nierozróżnialnych.

□

Symetryczność:

Weźmy dowolne dwa zbiory sekwencji: *A* i *B* oraz ustalmy dodatni parametr *k*. Sprawdźmy, czy

$$D(A,B) = D(B,A).$$

Zauważmy, że indeks Jaccarda jest symetryczny, to znaczy:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|B \cap A|}{|B \cup A|} = J(B,A).$$

Stąd też, dla dowolnych dwóch zbiorów *A* i *B* oraz dodatniego parametru *k*:

$$D(A,B) = -\frac{1}{k} ln \frac{2J(A,B)}{1+J(A,B)} = -\frac{1}{k} ln \frac{2J(B,A)}{1+J(B,A)} = D(B,A).$$

□

Nierówność trójkąta:

Weźmy dowolne trzy zbiory sekwencji: *A, B* i *C,* parami przecinające się, oraz ustalmy dodatni parametr *k*.

Sprawdźmy, czy dla tych zbiorów zachodzi nierówność:

$$\left(-\frac{1}{k}\right) ln \frac{2J(A,B)}{1+J(A,B)} + \left(-\frac{1}{k}\right) ln \frac{2J(B,C)}{1+J(B,C)} \geq \left(-\frac{1}{k}\right) ln \frac{2J(A,C)}{1+J(A,C)}.$$

Po przekształceniu (pamiętając o założeniu, że przekroje zbiorów są niepuste, w szczególności: $J(A,C) \neq 0$) jest to równoważne nierówności:

$$ln\left(\frac{2J(A,B)}{1+J(A,B)} * \frac{2J(B,C)}{1+J(B,C)} * \frac{1+J(A,C)}{2J(A,C)}\right) \leq 0.$$

Ponieważ logarytm naturalny jest funkcją monotoniczną rosnącą, równoważną nierównością dla tej powyższej jest:

$$\frac{2J(A,B) * 2J(B,C) * [1 + J(A,C)]}{[1 + J(A,B)] * [1 + J(B,C)] * 2J(A,C)} \leq 1.$$

Z definicji indeksu Jaccarda oraz z założenia $J(A,C) \neq 0$ mianownik ułamka z lewej strony jest ściśle dodatni, zatem wystarczy sprawdzić, że zachodzi:

$$2J(A,B) * 2J(B,C) * [1 + J(A,C)] \leq [1 + J(A,B)] * [1 + J(B,C)] * [2J(A,C)].$$

Zatem:

$$2J(A,B) * J(B,C) + J(A,B) * J(B,C) * J(A,C) \leq J(A,C) + J(A,B) * J(A,C) + J(B,C) * J(A,C),$$

$$2J(A,B) * J(B,C) \leq J(A,C)[1 + J(A,B) + J(B,C) - J(A,B) * J(B,C)].$$

Korzystając z faktu, że $1 + x + y - xy = (1 + x)(1 - y) + 2y$:

$$2J(A,B) * J(B,C) \leq J(A,C)[1 + J(A,B)][1 - J(B,C)] + 2J(B,C).$$

Odejmując obustronnie $2J(B,C)$ otrzymujemy:

$$2J(B,C) * [J(A,B) - 1] \leq J(A,C)[1 + J(A,B)][1 - J(B,C)],$$

co musi zachodzić, ponieważ po lewej stronie mamy iloczyn liczby dodatniej i liczby co najwyżej równej zero, zatem wartość jest niedodatnia, natomiast po prawej stronie równania mamy iloczyn liczb nieujemnych, zatem jest to wartość nieujemna.

□

3. Cel pracy

Celem tej pracy jest zaproponowanie nowego sposobu analizy danych transkryptomicznych, opartego na odległości. Dla danych pozyskanych w wyniku sekwencjonowania transkryptomu stosuje się głównie metody oparte na porównaniu ilościowym - np. poprzez kontrasty w metodach Deseq2 (Love i in., 2014) lub ALDEx2 (DOI: 10.18129/B9.bioc.ALDEx2), badanie poziomu ekspresji konkretnych genów, liczenie ekspresji na poziomie izoform itd. - natomiast dla analizy jakościowej wykorzystywane metody bazują na przykład na prostych porównaniach za pomocą narzędzia blast lub szukaniu polimorfizmów. Pojęcie odpowiednio zdefiniowanej odległości pozwala na uwzględnienie aspektu jakościowego w analizie zbiorów danych transkryptomicznych i wyrażeniu różnic pomiędzy nimi w sposób ilościowy. Wystarczy znaleźć odległość, której funkcja zależy od jakościowych różnic między zbiorami sekwencjami – im więcej różnic między daną parą zbiorów, tym wyższą liczbę przyporządkuje im funkcja. Taką odległością jest odległość Mash. Biorąc pod uwagę powyższe, sformułowany został główny cel tej pracy:

Pierwszy cel: opracowanie nowej metody analizy wyników doświadczeń wieloczynnikowych, w których uzyskano zbiory sekwencji transkryptomicznych (RNA).

Analizując dane uzyskane w eksperymentach biologicznych stosujemy funkcje i obliczenia mające na celu uporządkowanie wyników, znalezienie pewnych zależności, oszacowań czy tendencji. Pracujemy na liczbach, jednak efekty tej pracy ostatecznie potrzebują interpretacji związanej z obiektami biologicznymi, dla których zostały pozyskane. To prowadzi nas do drugiego celu:

Drugi cel: znalezienie sposobu na uwzględnienie w wyżej wymienionej metodzie aspektu biologicznego.

Patrząc w drugą stronę: eksperymenty wieloczynnikowe są właściwym sposobem badania różnic ilościowych między transkryptomami. Miara odległości umożliwia ilościowe opisanie różnic jakościowych między transkryptomami. Sformułowany zatem został cel trzeci:

Trzeci cel: sprawdzenie, czy uwzględnienie odległości Mash w porównawczej analizie ilościowej może poprawić wyniki wnioskowania.

4. Metody

W tym rozdziale opisane zostaną metody i narzędzia wykorzystane w pracy badawczej.

4.1. Przetwarzanie danych sekwencyjnych

4.1.1. Kontrola jakości danych

Kontrola jakości danych polega na sprawdzeniu poprawności, liczby oraz pewności odczytów. Format plików wynikowych sekwencjonowania to FASTQ: każdy odczyt zapisany jest w czterech wierszach, z czego pierwszy stanowi identyfikator odczytu i zaczyna się od znaku „@”, drugi jest zapisem odczytanych nukleotydów w postaci ciągu znaków składającego się z „A”, „C”, „G” i „T” oraz „N” („N” oznacza nierozpoznany nukleotyd lub niepewny odczyt), trzeci wiersz to jeden znak „+”, a czwarty jest ciągiem symboli oznaczających jakość odczytu, określoną w skali Phred, zakodowaną przy wykorzystaniu symboli ASCII. Jakość w skali Phred wyliczana jest wzorem $Q = -10\log_{10}p$, gdzie p oznacza prawdopodobieństwo błędnego odczytu. Dla analizowanych danych sprawdzono liczby odczytów w każdym zbiorze danych, statystyki dotyczące ich jakości (np. rozkład długości odczytów, liczba duplikatów, sekwencje, dla których wystąpiła nadreprezentacja), a także zawartość par GC (ang. GC-content, zawartość guaniny i cytozyny w sekwencjach, pozwala m.in. sprawdzić, czy nie występują pewne anomalie w proporcjach nukleotydów w odczytach), co wykonano korzystając z programu *FastQC* (*FastQC* v0.11.4; Andrews, 2010). Usunięcie fragmentów końcowych odczytów, które nie zostały rozpoznane (N) lub ich odczyty były słabej jakości (poniżej określonego poziomu dokładności) pomaga osiągnąć bardziej pewne wyniki w następnych krokach analizy. Należy również usunąć adaptery, czyli sekwencje nie będące transkryptami, a fragmentami przygotowanymi syntetycznie w celu umożliwienia sekwencjonowania. Warto również odrzucić odczyty, które po procesie obróbki okazały się za krótkie. Do wykonania tych zadań posłużył program *AdapterRemoval* (AdapterRemoval 2.1.7; Schubert i in., 2016).

4.1.2. Składanie odczytów

W celu otrzymania długich sekwencji transkryptów, tzw. kontigów (ang. contig; pojęcie to ma parę znaczeń, między innymi określa zbiór nakładających się fragmentów DNA, tu natomiast jest używane w kontekście odczytów tychże fragmentów połączonych zgodnie z miejscami nakładania się w konsensusowy odczyt jednego, dłuższego fragmentu) lub skafoldów (ang. scaffold, to znaczy zbiór kontigów wraz z przerwami między nimi o znanej, określonej długości), dane zostały poddane przetwarzaniu narzędziem *SOAPdenovo2-master* (wersja programu: SOAPdenovo-63mer 2.04-r241). Jako dane wejściowe program przyjmuje zbiory sekwencji w formacie FASTA, FASTQ lub BAM oraz stworzone dla nich pliki konfiguracyjne, zawierające niezbędne informacje o odczytach i zsekwencjonowanej bibliotece fragmentów. Program SOAPdenovo2-master wykorzystuje do składania sekwencji metodę opartą na grafach deBruijna (Luo i in., 2012), czyli grafach, których wierzchołki reprezentują *k*-mery znalezione w danym zbiorze sekwencji, a krawędzie łączą tylko te pary wierzchołków, dla których znaleziona została „nakładka” długości *k*-1, to znaczy *k*-1-mer znajdujący się na końcu sekwencji odpowiadającej jednemu wierzchołkowi i na początku sekwencji odpowiadającej drugiemu z nich.

4.2. Liczenie odległości Mash

Do obliczenia odległości Mash między wszystkimi parami zbiorów danych posłużył program o tej samej nazwie (Ondov i in., 2016). Jako dane wejściowe program przyjmuje pliki w formatach FASTQ lub FASTA, w których zapisane są nukleotydowe dane sekwencyjne oraz parametry takie jak długość k-merów, które mają zostać użyte w obliczeniach i liczba haszy w szkicu. Domyślnie algorytm Mash korzysta z alfabetu {A,C,G,T}, chociaż można ustawić inny zbiór znaków (na przykład zbiór liter oznaczających aminokwasy). Przy najprostszym stosowaniu oprogramowania wystarczy wywołać funkcję podając pary plików danych, dla których odległość ma zostać policzona. Jako wynik otrzymujemy tabelę z identyfikatorami obu plików, odległością Mash, liczbą haszy (reprezentujących k-mery), które były wspólne dla zbiorów danych oraz wartość prawdopodobieństwa przypadkowego otrzymania takiej lub większej liczby haszy. W przypadku dużej liczby zbiorów danych istnieje opcja stworzenia szkicu przed liczeniem odległości, co przyspiesza obliczenia. W tej pracy nie zdecydowano się na to z powodu i tak krótkiego czasu wykonywania działań przez algorytm.

4.3. Analiza współrzędnych głównych

Na macierzy odległości proponujemy przeprowadzić analizę współrzędnych głównych (Principal Coordinate Analysis, PCoA; Zuur i in., 2007). Metoda ta pozwala na przedstawienie graficzne wielowymiarowych danych opisanych za pomocą macierzy odległości. Polega ona na znalezieniu niskowymiarowej przestrzeni, w której odległości euklidesowe między punktami dobrze przybliżają odległości w przestrzeni badanej. Numerycznie, polega ona na znalezieniu rozkładu własnego macierzy, która powstała przez transformację macierzy odległości w następujący sposób:

1) Najpierw mając macierz odległości D, której *ij*-ty element oznaczamy przez D_{ij}, obliczamy macierz A składającą się z elementów A_{ij} zdefiniowanych poniżej:

$$A_{ij} = -\frac{1}{2} D_{ij}^2.$$

2) Znając elementy macierzy A obliczamy macierz E według wzoru:

$$E_{ij} = A_{ij} - \overline{A}_{\iota} - \overline{A}_{J} + \overline{A},$$

gdzie $\overline{A}_{\iota}$ i $\overline{A}_{J}$ to średnie odpowiednio z elementów *i*-tego wiersza i *j*-tej kolumny macierzy A, natomiast $\bar{A}$ to średnia z wszystkich elementów macierzy. Wówczas wektory własne macierzy E odpowiadają współrzędnym głównym (ang. principal coordinates, PC). Przekształcenie działa w ten sposób, że największą część zmienności w przestrzeni oryginalnej przedstawia pierwsza współrzędna główna (PC1), potem druga współrzędna (PC2) i tak dalej. Korzystając z tych współrzędnych można przedstawić wielowymiarowe dane w przestrzeni np. dwuwymiarowej.

Analiza PCoA została wykonana z wykorzystaniem pakietu R (wersja 3.6.1; biblioteka: „ape").

Do wyników analizy PCoA załączone zostały wykresy Sheparda, to znaczy wykresy punktowe przedstawiające na osi rzędnych odległości Mash, a na osi odciętych odległości euklidesowe między punktami w niżej wymiarowej przestrzeni, których współrzędne uzyskano z analizy współrzędnych głównych. Rozważano jeden, dwa, trzy, cztery, pięć i sześć wymiarów, a do każdego przypadku

przeprowadzona została regresja liniowa odległości euklidesowych względem odległości Mash. Wyniki zgromadzono w odpowiednich tabelach.

Współczynnik determinacji pojawiający się w tabelach wynikowych jest wartością pozwalającą określić, jak dobrze dany model matematyczny (w naszym przypadku regresji liniowej) opisuje obserwacje. Jest on określony wzorem

$$R = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})},$$

gdzie $y_i, i = 1, \dots, n,$ oznacza obserwację zmiennej y, $\bar{y}$ oznacza średnią arytmetyczną wszystkich obserwacji, a $\hat{y}_i$ oznacza wartości zmiennej y wyliczone na podstawie modelu. Im wartość współczynnika determinacji jest bliższa 1, tym lepiej świadczy to o dopasowaniu modelu.

4.4. Analiza wariancji i kowariancji

W pracy wykonywana jest analiza wariancji dla danych sklasyfikowanych według poziomów trzech czynników doświadczalnych oraz według powtórzeń biologicznych. Oznaczmy przez y_{ijkr} obserwację dotyczącą poziomów i, j, k czynników - odpowiednio - pierwszego, drugiego i trzeciego, oraz powtórzenia biologicznego r. Wówczas model liniowy obserwacji będący podstawą analizy wariancji można zapisać w postaci:

$$y_{ijkr} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijkr}, \quad (1)$$

gdzie symbole α, β, γ oznaczają, odpowiednio, efekty główne (średnie) trzech czynników, symbole $(\alpha\beta)$, $(\alpha\gamma)$, $(\beta\gamma)$ efekty interakcyjne par czynników, symbol $(\alpha\beta\gamma)$ efekty interakcyjne drugiego rzędu trójki czynników, zaś indeksy dolne podają numery poziomów czynników, których te efekty dotyczą. Symbolem e_{ijkr} oznaczono niezależne błędy losowe o założonym rozkładzie normalnym z wartością średnią 0 i jednakowej wariancji. Analiza wariancji polega na wyznaczeniu sum kwadratów i średnich kwadratów odpowiadających wszystkim uwzględnionym efektom, a następnie na wyliczeniu ilorazów średnich kwadratów, czyli wartości statystyki F, które pozwalają na ocenę istotności statystycznej efektów (Elandt 1964).

W rozdz. 6.1 i 6.2 do analizy zmienności współrzędnych prób (transkryptomów otrzymanych w poszczególnych próbach) otrzymanych w analizie PCoA wykorzystywany jest model zredukowany w stosunku do modelu (1), nie zawierający efektów interakcji drugiego rzędu:

$$y_{ijkr} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + e_{ijkr}. \quad (2)$$

Inaczej mówiąc, efekty interakcyjne drugiego rzędu traktuje się jako powodowane przez losowe błędy doświadczalne (składowe błędów losowych e_{ijkr}). Czynimy tak ze względu na często spotykane trudności w interpretacji efektów interakcji drugiego rzędu oraz chęć zwiększenia liczby stopni swobody dla estymacji wariancji błędów doświadczalnych.

W rozdz. 6.3 wykonywana jest analiza ekspresji genów za pomocą metody zaimplementowanej w funkcji Deseq2 w pakiecie R (Love i in., 2014). Funkcja ta co prawda nie wykonuje standardowej analizy wariancji dla zmiennej opisującej poziom ekspresji genu w poszczególnych próbach, jednak również posługuje się modelem liniowym. Dla doświadczenia trójczynnikowego mogą to być modele

(1) lub (2), jednakże autorzy metody Deseq2 rekomendują użycie modelu o innej parametryzacji, postaci

$$y_{ijkr} = m_{ijk} + e_{ijkr}, \tag{3}$$

gdzie m_{ijk} oznacza wartość oczekiwaną modelowanej zmiennej obserwowanej w próbie odpowiadającej kombinacji poziomów czynników (*i*,*j*,*k*). Metoda przewiduje estymację tych wartości oczekiwanych oraz testowanie istotności porównań (kontrastów) typu ($m_{ijk} - m_{i',j',k'}$), co pozwala na ocenę różnic ekspresji genów dla różnych wariantów doświadczalnych.

W rozdz. 6.3 zastosowana została też modyfikacja metody Deseq2 wykorzystująca zamiast modelu liniowego doświadczenia wieloczynnikowego odpowiadającego analizie wariancji, jego modyfikację do postaci modelu analizy kowariancji. Jeżeli dostępne są obserwacje dodatkowych zmiennych objaśniających (towarzyszących) $z^p, p = 1,2 \dots P$, które mogą wyjaśniać różnice pomiędzy wariantami pod względem zmiennej modelowanej, to model analizy kowariancji ma postać

$$y_{ijkr} = m_{ijk} + \sum_{p=1}^{P} r^p z_{ijkr}^p + e_{ijkr}, \tag{4}$$

gdzie z_{ijkr}^p oznacza obserwację *p*-tej zmiennej objaśniającej dla próby (*i*,*j*,*k*,*r*), zaś r^p oznacza współczynnik regresji zmiennej modelowanej (zależnej) względem *p*-tej zmiennej objaśniającej. W rozdz. 6.3 rolę zmiennych objaśniających pełnią współrzędne prób wyznaczone w analizie współrzędnych głównych macierzy odległości Mash pomiędzy transkryptomami obserwowanymi w poszczególnych próbach.

W pracy wykorzystywana jest również wielowymiarowa analiza wariancji (Morrison, 1967). Podobnie jak w przypadku zastosowanej przez nas jednowymiarowej analizy, przeprowadzana jest ona dla danych sklasyfikowanych według poziomów trzech czynników doświadczalnych oraz według powtórzeń biologicznych, jednak tym razem rolę zmiennych objaśnianych pełnią zestawy czterech pierwszych współrzędnych prób, wyznaczone w analizie współrzędnych głównych dla macierzy odległości Mash.

Analogicznie do analizy wariancji, model zostanie zredukowany poprzez pominięcie efektów interakcji drugiego rzędu. Model obserwacji będzie wyglądał jak we wzorze (2), z tą różnicą, że y_{ijkr} będzie wektorem wierszowym $[y_1, y_2, y_3, y_4]_{ijkr}$, podobnie czterowymiarowymi wektorami będą: μ, α_i, β_j, γ_k, $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$, $(\alpha\beta\gamma)_{ijk}$ oraz e_{ijkr}.

Do dalszych rozważań przyjmijmy, że czynniki pierwszy i drugi (o efektach α i β) mają po dwa poziomy, natomiast czynnik trzeci (o efekcie γ) ma trzy poziomy. Ogólny zapis macierzowy modelu liniowego to:

$$Y = A\xi + \varepsilon, \tag{5}$$

gdzie dla naszego konkretnego przypadku:

Y to macierz o wymiarach 36×4, której każda kolumna to kolejno pierwsze, drugie, trzecie i czwarte współrzędne wszystkich prób. To dla tych współrzędnych poszukujemy modelu, który by je opisywał

w zależności od czynników i ich kombinacji. Załóżmy, że wiadomo, że w macierzy *Y* w *m*-tym rzędzie mamy obserwację ustalonych poziomów *i, j, k* odpowiednio czynników α, β oraz γ. Przyjmijmy taką kolejność elementów, że w *n*-tej kolumnie i *m*-tym wierszu macierzy *Y* znajduje się element:

$$y_{2-mod\left(\left\lceil\frac{m}{6}\right\rceil,2\right),\ 2-mod\left(\left\lceil\frac{m}{3}\right\rceil,2\right),\ \left\lceil\frac{m}{12}\right\rceil,\ mod(m-1,3)+1,\ n} \tag{6}$$

dla $m \in \{1, \dots, 36\}, n \in \{1,2,3,4\}$, gdzie mod(*a,b*) to reszta z dzielenia *a* przez *b*, a $\left\lceil\frac{a}{b}\right\rceil$ to wynik dzielenia *a* przez *b* zaokrąglony do całości w górę. To znaczy, że w powyższym zapisie mamy: $i = 2 - mod\left(\left\lceil\frac{m}{6}\right\rceil, 2\right), j = 2 - mod\left(\left\lceil\frac{m}{3}\right\rceil, 2\right), k = \left\lceil\frac{m}{12}\right\rceil$, a wartość $mod(m-1,3)+1$ to numer powtórzenia *r*.

ξ to macierz parametrów 24×4. Przy przyjętych założeniach *n*-ta kolumna tej macierzy dla *n*=1,2,3,4 będzie miała postać:

$$[\alpha_{1,n}, \alpha_{2,n}, \beta_{1,n}, \beta_{2,n}, \gamma_{1,n}, \gamma_{2,n}, \gamma_{3,n}, (\alpha\beta)_{1,1,n}, (\alpha\beta)_{1,2,n}, (\alpha\beta)_{2,1,n}, (\alpha\beta)_{2,2,n}, (\alpha\gamma)_{1,1,n}, \tag{7}$$
$$(\alpha\gamma)_{2,1,n}, (\alpha\gamma)_{1,2,n}, (\alpha\gamma)_{2,2,n}, (\alpha\gamma)_{1,3,n}, (\alpha\gamma)_{2,3,n}, (\beta\gamma)_{1,1,n}, (\beta\gamma)_{2,1,n}, (\beta\gamma)_{1,2,n},$$
$$(\beta\gamma)_{2,2,n}, (\beta\gamma)_{1,3,n}, (\beta\gamma)_{2,3,n}, \mu_n\ ]'.$$

A to macierz 36×24 składająca się z zer i jedynek (macierz indykatorów), skonstruowana w taki sposób, żeby wskazywać współczynniki odpowiadające czynnikom oraz ich interakcjom na zaobserwowanych poziomach. Przy powyższych założeniach w *t*-tym rzędzie jedynki są na miejscach: *i*-tym, $(2 + j)$-tym, $(4 + k)$-tym, $(7 + 2(i-1) + j)$-tym, $(11 + 2(k-1) + i)$-tym, $(17 + 2(k-1) + j)$-tym i 24-tym.

ε to macierz błędów, o wymiarze takim samym jak Y, to znaczy 36×4.

4.5. Porównawcza analiza ekspresji genów

4.5.1. Mapowanie odczytów

Aby przypisać odczytom NGS sekwencji genetycznych odpowiadające im geny czy produkty, należy przyrównać je do znanych nam już sekwencji - tak zwanych genomów lub transkryptomów referencyjnych. Kiedy badamy organizm, który został wcześniej w pełni zsekwencjonowany, możemy skorzystać z referencji dla danego organizmu znajdujących się w bazach danych, natomiast w przypadku gatunków, których genom nie jest jeszcze znany, najczęściej poszukujemy najbliżej spokrewnionych gatunków. Mając taką referencję można wykonać mapowanie, które w tym przypadku zostało przeprowadzone za pomocą narzędzia TopHat (ver. 2.1.1.). Istnieje również opcja opracowania referencyjnego genomu lub transkryptomu przy pomocy długich odczytów (sekwencjonowanie III generacji), a następnie mapowania sekwencji w tak zbudowanej referencji.

4.5.2. Analiza różnicowa ekspresji genów

W rozdziale 6.3 do porównania poziomów ekspresji genów między wybranymi zbiorami danych posłużył nam pakiet DESeq2 (Love i in., 2014) obsługiwany przez *R*.

4.6. Ontologie genów

Aby wprowadzić kontekst funkcjonalny do analizy odległości między sekwencjami, wykorzystaliśmy informacje pozyskane z ontologii genów Gene Ontology, dokładniej z GOSlim GOA z bazy danych Ensembl Plants.

Gene Ontology (Ashburner i in., 2000; The Gene Ontology Consortium 2023) to projekt bioinformatyczny, którego celem jest usystematyzowanie informacji na temat genów i ich produktów oraz umożliwienie rozbudowywania i aktualizowania ich w miarę rozwoju wiedzy na ten temat. Konstrukcja ontologii i uporządkowane słownictwo pozwalają zarówno na zachowanie niezależności gatunkowej przy jej rozbudowywaniu, jak i na proste przeszukiwanie w celu znalezienia interesujących nas danych. Strukturę ontologii można przedstawić jako graf. Jego wierzchołkami są klasy zwane również terminami (ang. „classes”/”terms”), a krawędziami - zależności: bycie podklasą lub nadklasą. Istnieją trzy aspekty, względem których rozpatrywane są geny lub ich produkty: funkcja molekularna (ang. „Molecular Function”), składnik komórki (ang. „Cellular Component”) oraz proces biologiczny (ang. „Biological Process”). Te trzy domeny wyznaczają trzy rozdzielne spójne grafy. W funkcjach molekularnych zawarte są działania wykonywane przez produkty genów na poziomie molekularnym. Aby uniknąć niejednoznaczności, w nazwach terminów funkcji molekularnych pojawia się słowo „activity” (aktywność). Składnik komórki to domena odnosząca się do anatomii komórkowej. Określa miejsce, w którym dany produkt pełni funkcję. Z kolei w procesie biologicznym zawarte są terminy odnoszące się do procesów realizowanych przez wiele działań molekularnych (na przykład: transdukcja sygnału, naprawa DNA).

GO Slims to uogólnione ontologie tworzone na potrzeby badaczy, w celu szybszej selekcji i łatwiejszej obserwacji interesujących ich zjawisk.

Do przedstawienia danych w formie czytelnych wykresów wykorzystano narzędzie CytoScape (CytoScape v. 3.8.1) - program umożliwiający budowanie oraz wizualizację grafów.

5. Opis danych doświadczalnych

Do testowania opracowanej metody analizy wykorzystano dane eksperymentalne z badań reakcji jęczmienia jarego (*Hordeum vulgare* L.) na podwyższoną temperaturę środowiska (Mikołajczak i in., 2022). Pod uwagę brane były trzy genotypy:

- Bowman (BW, dwurzędowa odmiana jęczmienia jarego)

oraz jego dwie niemal izogeniczne linie (NILs):

- BW827, posiadająca mutację sdw1.a;
- BW828, posiadająca mutację sdw1.d.

Niemal izogeniczne linie BW827 i BW828 zostały otrzymane przez wielokrotne krzyżowanie wsteczne mutantów z odmianą Bowman. U BW827, posiadającego mutację sdw1.a, stwierdzono całkowitą delecję genu *HvGA20ox2*. BW828 posiada 7-nukleotydową delecję w tym genie, która powoduje przesunięcie ramki odczytu i przedwczesne zakończenie translacji, jednak nadal może on częściowo pełnić funkcję oksydazy GA-20. Zatem mutacje te powodują, odpowiednio, niemal całkowite i częściowe ograniczenie syntezy oksydazy GA-20, która pełni kluczową rolę w syntezie kwasu giberelinowego, który z kolei stymuluje wzrost i rozwój roślin (Xu i in. 2017).

Jęczmień rósł w doniczkach, w kontrolowanych warunkach. Grupa kontrolna (OT) uprawiana była w temperaturze 8°C w nocy i 16°C za dnia w okresie między siewem a końcem krzewienia, natomiast między końcem krzewienia a dojrzałością - w temperaturze 12°C w nocy i 20°C za dnia. Grupa traktowana wysoką temperaturą (HT) w okresie między wysiewem a końcem krzewienia trzymana była w temperaturze 28°C, natomiast między końcem krzewienia a dojrzałością w takich samych warunkach jak grupa kontrolna, czyli w temperaturze 12°C w nocy i 20°C za dnia. Materiał pobrano w dwóch punktach czasowych: pierwszy dzień krzewienia (23-26 dzień w skali BBCH) i 10 dni później, przy czym wykonano po trzy powtórzenia biologiczne. Z próbek pochodzących z tzw. koron roślin jęczmienia (korona to pierwszy węzeł powyżej ziarna, część rośliny łącząca korzeń i pęd, która powstaje w czasie rozwoju siewki; wytwarza pąki, z których rozwijają się pędy, korzenie przybyszowe, kłącza i rozłogi, a także pełni rolę magazynu rezerwy węglowodanów mającego na celu wsparcie wzrostu nowych organów; merystemy w koronie orpowiadają za regenerację pędów i korzeni rośliny następującą po stresie) wyekstrahowano RNA przy użyciu TRI Reagent® RT (Molecular Research Center, Inc., Cincinnati, OH, USA), a następnie oczyszczono otrzymany RNA przy pomocy DNazy I. W kolejnym kroku sprawdzono ilość i jakość przy użyciu spektrofotometru NanoDrop 2000 (Thermo Fisher Scientific) oraz za pomocą stacji elektroforezy Experion^{TM} potwierdzono, że wskaźnik integralności RNA (RIN) jest wystarczający do przeprowadzenia sekwencjonowania. Konstrukcję bibliotek cDNA (TruSeq Stranded mRNA LT Sample Prep Kit) oraz sekwencjonowanie przeprowadziła firma Macrogen Inc. za pomocą platformy Illumina NovaSeq 6000, przy parametrach: 2 x 101 par zasad, sparowaniu końców. Wykorzystane tutaj dane są dostępne w bazie danych ArrayExpress pod numerem akcesyjnym E-MTAB-10789.

6. Wyniki

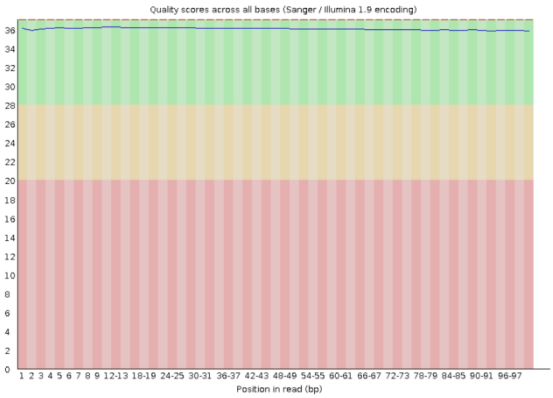
6.1. Analiza podobieństwa transkryptomów w doświadczeniach wieloczynnikowych

W wyniku sekwencjonowania bibliotek RNA metodą Illumina otrzymano 36 zbiorów danych (Tabela 6.1). Nazwy prób widoczne w pierwszej kolumnie tabeli będą używane również w następnych rozdziałach. Pierwszy człon każdej nazwy próby oznacza genotyp, drugi człon odnosi się do traktowania roślin (oznaczenia: HT – wysoka temperatura, OT – optymalna temperatura), trzeci człon oznacza punkt czasowy (1d – pierwszy dzień krzewienia, 10d – dziesięć dni później), a czwarty to numer powtórzenia (R1, R2 i R3). Uzyskano od 17,9 do 26,0 M par odczytów na próbkę. Kontrola danych nie wykazała nieprawidłowości w zakresie jakości odczytów i zawartości GC (Ryc. 6.1)

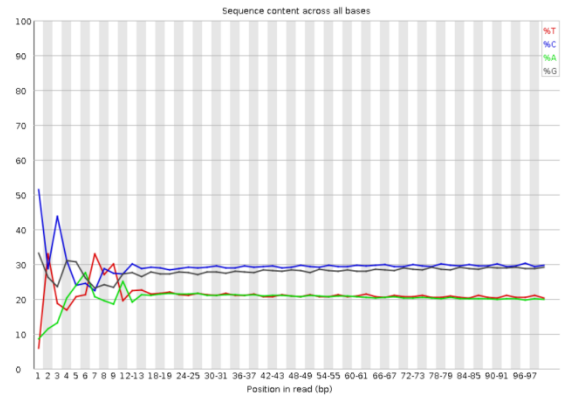
per base sequence quality, _1



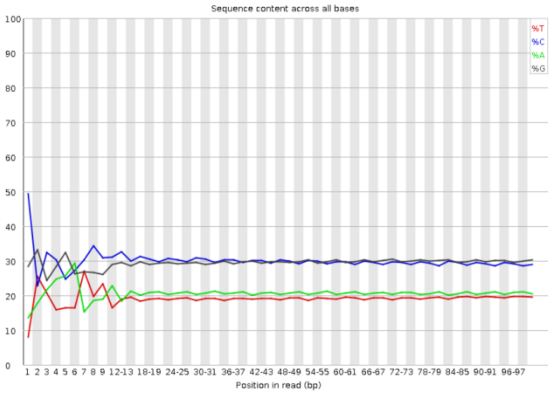
per base sequence quality, _2



GC sequence content, _1



GC sequence content, _2



Ryc. 6.1. Kontrola jakości odczytów lewych (_1) i prawych (_2) za pomocą programu FastQC dla próby BW828OT10dR1 jako przykład wyników kontroli jakości danych. Jakość sekwencji była zadowalająca, zawartość i proporcje GC również nie wskazywały na błędy sekwencjonowania, które mogłyby niepokoić.

Dane zostały poddane obróbce, w tym usuwaniu adapterów, za pomocą programu *AdapterRemoval*. Jako minimalną akceptowalną jakość przyjęto 20 (parametr: --minquality 20), z końców sekwencji usunięto fragmenty, które nie zostały odczytane (N) lub ich odczyty były słabej jakości (parametry: --trimns --trimqualities), a odczyty, które po tym procesie okazały się być krótsze niż 50 nukleotydów, zostały usunięte (parametr: --minlength 50). Obróbka i filtrowanie ze względu na jakość dostarczyły zbiorów danych o liczbie odczytów stanowiących co najmniej 99,94% liczby danych surowych (Tabela 6.1).

Składanie odczytów w programie *SOAPdenovo2* wykonano stosując długość k-merów równą 63 (K = 63). Ponieważ odczyty uzyskane w wyniku sekwencjonowania miały długość 101 nukleotydów, jako maksymalną długość odczytu ustawiono 101 (max_rd_len = 101), a jako średnią długość insertów ustawiono 180 (avg_ins = 180), zakładając średnią długość fragmentów ok. 380 nukleotydów. Pod uwagę brane były zarówno kontigi, jak i skafoldy (asm_flags = 3), co znaczy, że jako wynik otrzymaliśmy dla zbiorów danych osobno kontigi oraz skafoldy. W Tabeli 6.2 podana została liczba kontigów i skafoldów uzyskanych dla poszczególnych prób oraz średnia długość, mediana i maksymalna długość dla jednych i drugich.

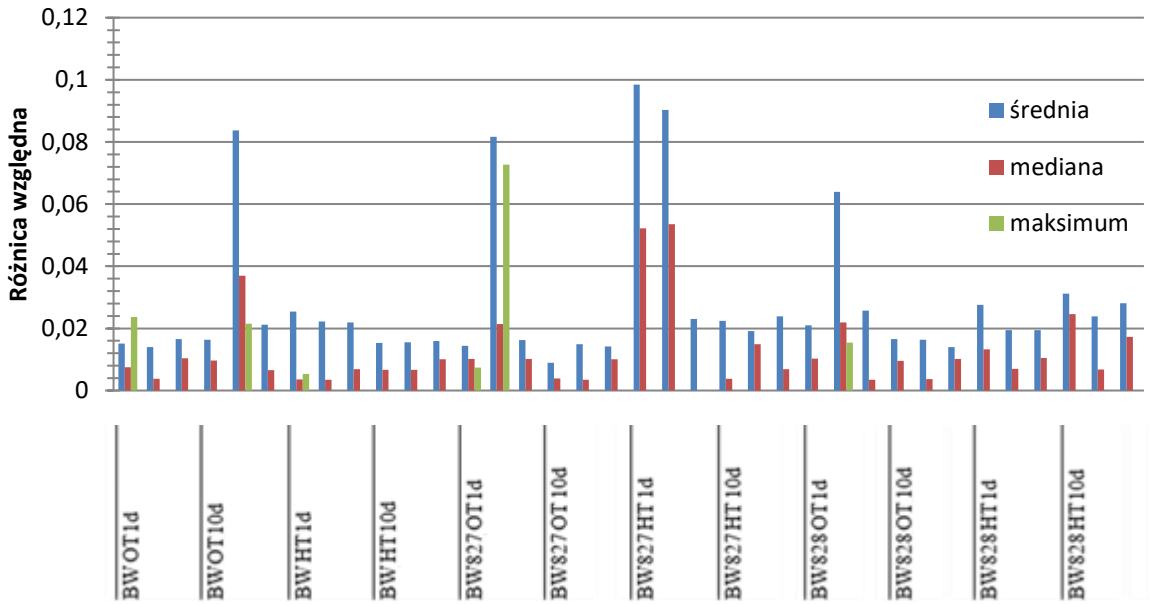
Tabela 6.1. Charakterystyki danych NGS przed i po obróbce jakościowej

Próba	Liczba par odczytów w danych surowych	Liczba par odczytów po obróbce jakościowej	Procent danych pozostawionych po filtrowaniu ze względu na jakość
BWOT1dR1	25 102 660	25 101 713	99,996
BWOT1dR2	21 958 409	21 956 514	99,992
BWOT1dR3	18 942 686	18 942 090	99,997
BWOT10dR1	18 743 112	18 742 617	99,997
BWOT10dR2	24 879 568	24 878 836	99,997
BWOT10dR3	21 098 755	21 098 108	99,997
BWHT1dR1	23 584 479	23 583 623	99,996
BWHT1dR2	25 251 806	25 251 169	99,997
BWHT1dR3	22 277 752	22 277 076	99,997
BWHT10dR1	23 088 395	23 079 547	99,962
BWHT10dR2	19 029 517	19 029 040	99,997
BWHT10dR3	21 983 374	21 980 619	99,987
BW827OT1dR1	17 992 064	17 991 425	99,996
BW827OT1dR2	24 590 779	24 589 361	99,994
BW827OT1dR3	19 090 449	19 089 880	99,997
BW827OT10dR1	22 334 471	22 321 852	99,944
BW827OT10dR2	21 475 652	21 473 279	99,989
BW827OT10dR3	21 467 245	21 466 283	99,996
BW827HT1dR1	20 918 239	20 917 564	99,997
BW827HT1dR2	17 896 163	17 895 607	99,997
BW827HT1dR3	25 455 112	25 451 308	99,985
BW827HT10dR1	25 978 755	25 976 567	99,992
BW827HT10dR2	19 383 338	19 382 147	99,994
BW827HT10dR3	23 153 841	23 151 812	99,991
BW828OT1dR1	24 392 622	24 390 899	99,993
BW828OT1dR2	22 380 890	22 380 249	99,997
BW828OT1dR3	23 163 941	23 163 289	99,997
BW828OT10dR1	19 237 549	19 236 531	99,995
BW828OT10dR2	20 623 545	20 622 201	99,993
BW828OT10dR3	20 711 720	20 710 528	99,994
BW828HT1dR1	21 917 691	21 916 047	99,993
BW828HT1dR2	18 218 801	18 216 752	99,989
BW828HT1dR3	20 992 242	20 989 135	99,985
BW828HT10dR1	21 163 323	21 158 409	99,977
BW828HT10dR2	21 548 473	21 547 757	99,997
BW828HT10dR3	20 131 925	20 128 793	99,984

Tabela 6.2. Charakterystyka kontigów i skafoldów

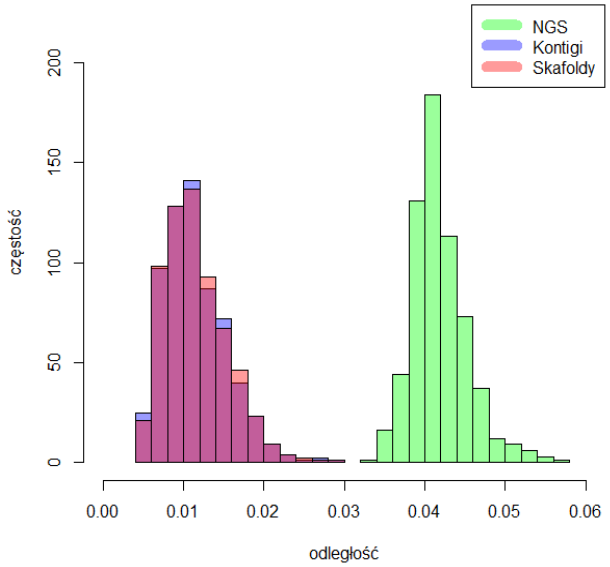
Próba	Kontigi				Skafoldy			
	Liczba	Średnia długość	Mediana	Maksimum	Liczba	Średnia długość	Mediana	Maksimum
BWOT1dR1	65 565	456	265	4 630	64 347	463	267	4 742
BWOT1dR2	57 102	422	263	4 372	56 139	428	264	4 372
BWOT1dR3	60 029	475	287	4 466	58 943	483	290	4 466
BWOT10dR1	58 702	541	309	7 544	57 569	550	312	7 544
BWOT10dR2	67 302	536	287	7 251	61 093	585	298	7 410
BWOT10dR3	62 789	554	301	8 053	61 315	566	303	8 053
BWHT1dR1	71 168	537	279	8 023	69 181	551	280	8 066
BWHT1dR2	70 144	529	284	9 549	68 370	541	285	9 549
BWHT1dR3	65 560	536	290	7 683	64 039	548	292	7 683
BWHT10dR1	64 868	514	298	7 537	63 747	522	300	7 537
BWHT10dR2	65 252	507	296	6 161	64 160	515	298	6 161
BWHT10dR3	66 631	495	295	7 025	65 416	503	298	7 025
BW827OT1dR1	50 764	480	292	4 416	49 954	487	295	4 449
BW827OT1dR2	68 053	495	275	6 989	62 041	539	281	7 537
BW827OT1dR3	58 196	486	291	4 622	57 122	494	294	4 622
BW827OT10dR1	61 321	445	260	7 091	60 624	449	261	7 091
BW827OT10dR2	54 979	462	285	7 025	54 111	469	286	7 025
BW827OT10dR3	60 162	488	294	7 029	59 111	495	297	7 029
BW827HT1dR1	126 307	403	254	7 548	112 188	447	268	7 548
BW827HT1dR2	56 829	594	318	11 500	51 215	653	336	11 500
BW827HT1dR3	70 950	509	258	11 454	69 169	521	258	11 454
BW827HT10dR1	68 871	524	266	11 976	67 080	536	267	11 976
BW827HT10dR2	92 501	410	265	5 682	90 230	418	269	5 682
BW827HT10dR3	70 425	531	288	11 440	68 621	544	290	11 440
BW828OT1dR1	60 604	466	288	4 420	59 182	476	291	4 420
BW828OT1dR2	60 232	439	268	4 468	55 933	469	274	4 538
BW828OT1dR3	70 228	531	284	10 267	68 200	545	285	10 267
BW828OT10dR1	55 948	536	311	7 378	54 913	545	314	7 378
BW828OT10dR2	62 074	481	273	7 006	60 974	489	274	7 006
BW828OT10dR3	58 305	493	291	6 519	57 300	500	294	6 519
BW828HT1dR1	64 248	565	297	11 487	62 239	581	301	11 487
BW828HT1dR2	61 647	504	283	7 245	60 383	514	285	7 245
BW828HT1dR3	58 665	455	283	4 151	57 270	464	286	4 151
BW828HT10dR1	64 018	529	318	7 626	61 792	546	326	7 626
BW828HT10dR2	70 969	531	292	8 055	69 052	544	294	8 055
BW828HT10dR3	85 226	449	285	5 762	82 465	462	290	5 762

Charakterystyki kontigów i skafoldów w większości przypadków nieznacznie różniły się między sobą. Pokazuje to Rycina 6.2, na której dla średniej, mediany i maksimum długości przedstawiono, o jaki ułamek większy był każdy z tych parametrów dla skafoldów niż dla kontigów. Wartości te nie przekroczyły 0,1 i w większości przypadków były mniejsze niż 0,03. Maksymalna długość zmieniła się tylko w sześciu przypadkach. Największa różnica wystąpiła dla próby BW827OT1dR2, gdzie maksimum dla skafoldów było o 7,3% większe niż dla kontigów, jednocześnie reszta różnic nie przekroczyła 2,3%. Różnice w medianach nie przekroczyły 5,4% i tylko dla trzech prób były to wartości powyżej 3%. Różnice średnich dla większości przypadków nie przekroczyły 3%, większą różnicę zaobserwowano dla sześciu prób: BWOT10dR2 – 8,4%, BW827OT1dR2 – 8,2%, BW827HT1dR1 – 9,8%, BW827HT1dR2 – 9,0%, BW828OT1dR2 – 6,4%, BW828HT10dR1 – 3,1%.



Ryc. 6.2. Różnice między charakterystykami skafoldów i kontigów podzielone przez wartość charakterystyk dla skafoldów. Każda trójka słupków prezentuje wyniki dla jednego powtórzenia biologicznego.

W następnym kroku policzone zostały odległości Mash między zbiorami danych, osobno dla każdego etapu przetwarzania danych: dla odczytów NGS po wstępnej obróbce danych, dla kontigów oraz dla skafoldów. Parametry użyte przy obliczeniach to parametry domyślne, czyli: długość k-merów równa 21 i wielkość szkicu (sketch) równa 1000. Wartości prawdopodobieństwa wyliczane przez program Mash były bardzo niskie - raportowane przez program jako 0, co wskazuje na wartości wystarczająco bliskie zeru - świadcząc o dużej wiarygodności wyników. Wynik obliczeń każdorazowo zebrano w macierzy odległości, która posłużyła do dalszej analizy. Otrzymano w ten sposób trzy macierze odległości. Dwie z nich - dla danych NGS i dla skafoldów - przedstawiono w Suplemencie, w Tabeli S1. Charakterystyki tych macierzy zostały podane na Rycinie 6.3 oraz w Tabeli 6.3 – dla kontigów i skafoldów różniły się one w niewielkim stopniu.



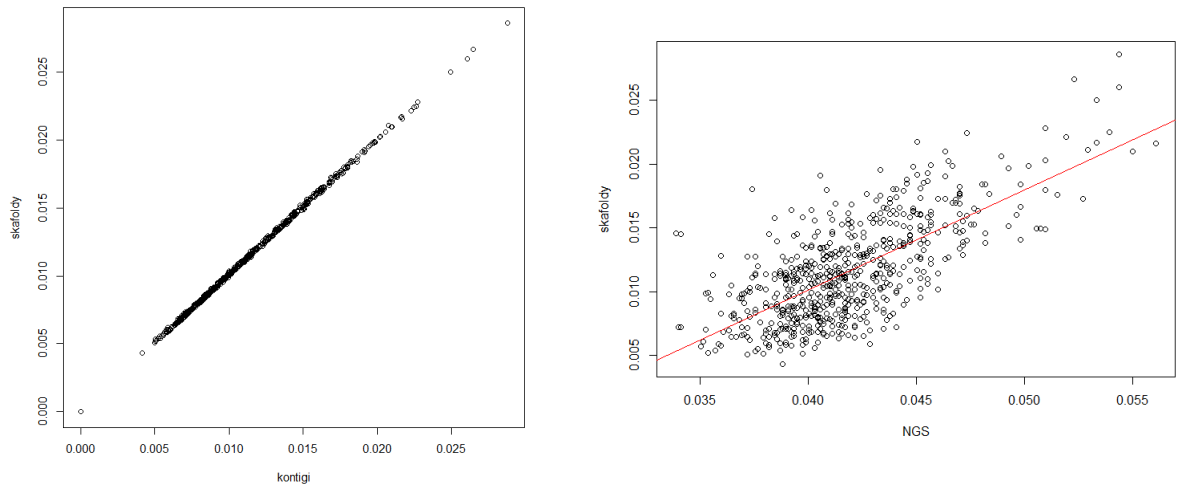
Ryc. 6.3. Częstość występowania danych odległości Mash dla trzech wariantów obliczeń: dla danych NGS, kontigów i skafoldów.

Tabela 6.3. Charakterystyki macierzy odległości

Wariant obliczeń	Minimum	Maksimum	Średnia
Dane NGS	0,0339	0,0561	0,0418
Kontigi	0,0041	0,0288	0,0115
Skafoldy	0,0043	0,0286	0,0116

Przeanalizujmy dokładniej zależności pomiędzy odległościami liczonymi dla zbiorów danych na różnych etapach składania (Ryc. 6.4). Patrząc na ilustrację 6.4.A widzimy, że odległości Mash dla zbiorów kontigów i zbiorów skafoldów były niemal identyczne. Dlatego też do dokładnych analiz postanowiliśmy wybrać tylko jeden z tych zbiorów: końcowy etap składania, czyli skafoldy. Patrząc z kolei na wykres dla NGS i skafoldów widzimy różnice w odległościach (6.4.B). Dla tego porównania policzona została regresja liniowa w R (funkcja lm). Współczynniki wyliczone tym sposobem to: -0,02125906 i 0,78436643. Na wykresie czerwoną linią została nakreślona prosta ilustrująca tę regresję: y = -0,021 + 0,784x. Do porównania odległości wyliczonych na różnych etapach analizy danych powrócimy w Dyskusji.

A. B.



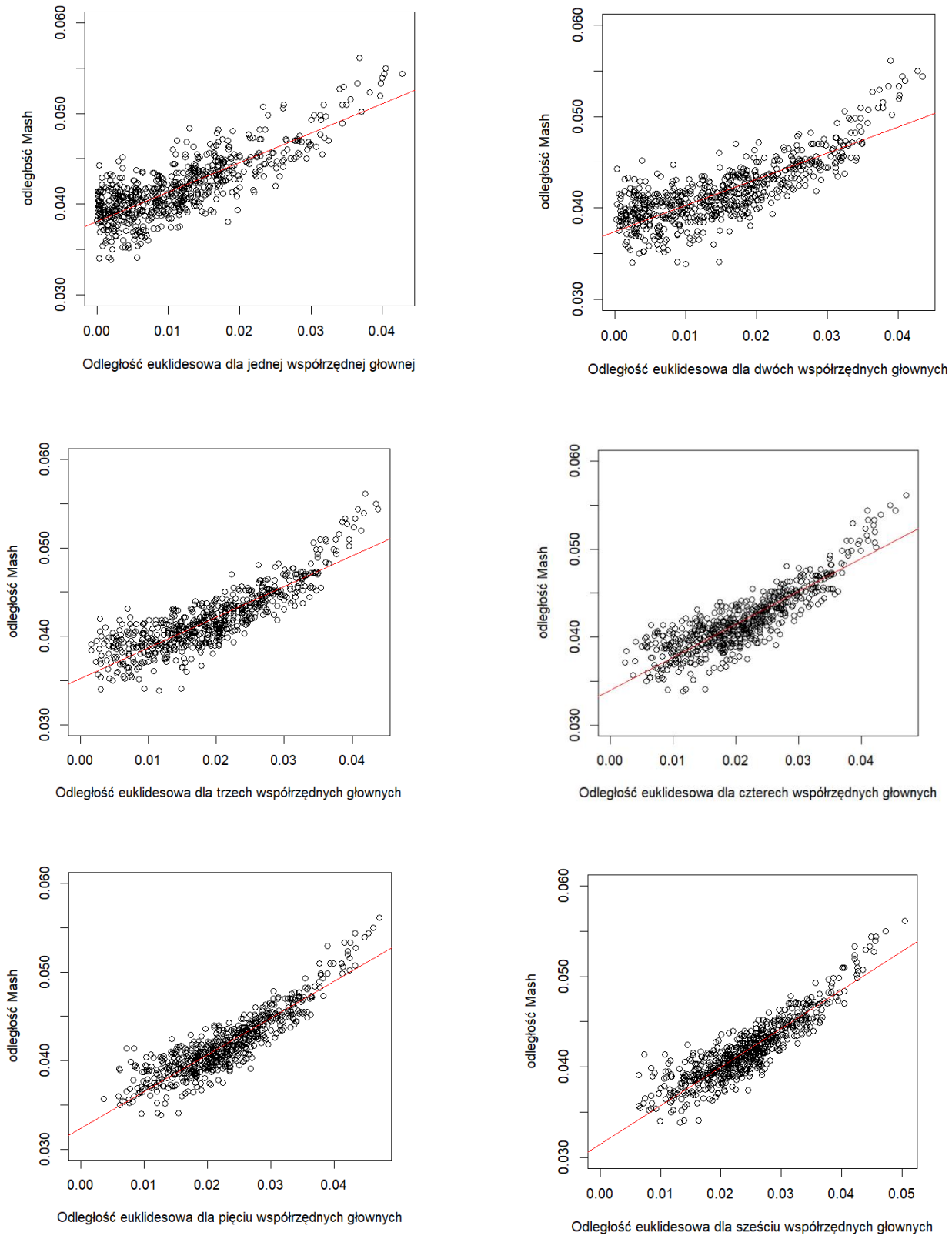
Ryc. 6.4. Wykres zależności pomiędzy odległościami Mash dla par zbiorów: A. kontigów i skafoldów, B. skafoldów i danych NGS;

6.1.1. Analiza odległości wykorzystująca dane NGS

Wykonano analizę PCoA dla macierzy odległości Mash liczonych na zbiorach danych NGS (po obróbce jakościowej) i wyliczono współrzędne prób w przestrzeni współrzędnych głównych (Tabela 6.4). Procent zmienności wyjaśnianej przez pierwsze pięć współrzędnych głównych wynosił kolejno:11,82%, 7,22%, 5,64%, 4,26% oraz 3,90%; całkowity procent zmienności dla 5 zmiennych: 32,84%.

Tabela 6.4. Wektory współrzędnych prób uzyskane z PCoA, pierwsze pięć współrzędnych głównych.

Próba	PC1	PC2	PC3	PC4	PC5
BWOT1dR1	-0,0121	0,0005	-0,0122	-0,0009	0,0063
BWOT1dR2	-0,0120	0,0093	-0,0006	-0,0088	0,0085
BWOT1dR3	-0,0082	0,0021	-0,0099	0,0020	0,0008
BWOT10dR1	-0,0029	-0,0005	-0,0059	0,0016	-0,0143
BWOT10dR2	0,0071	-0,0083	-0,0041	0,0013	-0,0137
BWOT10dR3	0,0021	-0,0083	-0,0107	0,0021	-0,0035
BWHT1dR1	0,0069	-0,0125	-0,0018	-0,0047	0,0032
BWHT1dR2	0,0047	-0,0098	-0,0050	-0,0014	0,0010
BWHT1dR3	0,0056	-0,0119	-0,0071	-0,0070	0,0038
BWHT10dR1	-0,0057	0,0016	0,0102	0,0037	0,0054
BWHT10dR2	0,0107	0,0079	-0,0074	0,0055	-0,0018
BWHT10dR3	-0,0021	-0,0004	0,0051	0,0077	0,0049
BW827OT1dR1	-0,0099	0,0019	-0,0053	-0,0051	0,0051
BW827OT1dR2	-0,0072	0,0013	-0,0010	0,0038	0,0057
BW827OT1dR3	-0,0066	0,0017	-0,0098	-0,0033	0,0032
BW827OT10dR1	-0,0114	0,0088	0,0053	-0,0030	-0,0047
BW827OT10dR2	-0,0145	0,0069	0,0007	0,0009	-0,0080
BW827OT10dR3	-0,0088	0,0020	0,0007	0,0042	-0,0090
BW827HT1dR1	0,0283	0,0146	-0,0047	-0,0110	-0,0058
BW827HT1dR2	0,0054	-0,0099	0,0026	-0,0085	-0,0042
BW827HT1dR3	0,0043	-0,0083	0,0124	-0,0017	0,0008
BW827HT10dR1	0,0023	-0,0048	0,0065	0,0086	0,0064
BW827HT10dR2	0,0179	0,0198	-0,0004	0,0036	0,0012
BW827HT10dR3	0,0162	0,0052	0,0015	0,0064	0,0041
BW828OT1dR1	-0,0117	0,0054	0,0030	-0,0003	0,0063
BW828OT1dR2	-0,0115	0,0096	-0,0099	0,0004	0,0007
BW828OT1dR3	0,0050	-0,0126	-0,0097	0,0108	0,0043
BW828OT10dR1	-0,0034	0,0004	0,0050	0,0056	-0,0097
BW828OT10dR2	-0,0061	-0,0013	0,0065	0,0003	-0,0036
BW828OT10dR3	-0,0085	0,0019	0,0109	0,0106	-0,0053
BW828HT1dR1	0,0055	-0,0097	0,0105	-0,0042	0,0004
BW828HT1dR2	0,0000	-0,0029	0,0073	-0,0117	-0,0007
BW828HT1dR3	-0,0062	0,0002	0,0046	-0,0133	-0,0005
BW828HT10dR1	0,0002	-0,0005	0,0089	-0,0030	0,0020
BW828HT10dR2	0,0065	-0,0095	-0,0014	0,0061	0,0045
BW828HT10dR3	0,0199	0,0100	0,0053	0,0026	0,0064



Ryc. 6.5. Wykresy Sheparda dla surowych danych, kolejno dla jednej, dwóch, trzech, czterech, pięciu i sześciu pierwszych współrzędnych głównych.

Rycina powyżej (Ryc. 6.5) przedstawia relację odległości Mash między zbiorami danych i odpowiadających im odległości euklidesowych wyliczonych dla jednej, dwóch, trzech, czterech, pięciu i sześciu współrzędnych uzyskanych w wyniku analizy PCoA. Proste na wykresach wyznaczone zostały przy użyciu regresji liniowej, której wyniki zostały zawarte w Tabeli 6.5.

Tabela 6.5. Regresja liniowa policzona dla odległości Mash i odległości euklidesowych.

	Współczynniki regresji i ich błędy standardowe (regresja liniowa; funkcja postaci: $y = ax + b$)				Współczynnik determinacji	Wartość P dla statystyki F
Liczba wymiarów	*a*	błąd standardowy	*b*	błąd standardowy		
1	0,2043079	0,0278831	0,0406189	0,0002139	0,07729	7,245e-13
2	0,2775980	0,0288460	0,0393810	0,0002870	0,12710	< 2,2e-16
3	0,2526609	0,0306355	0,0392983	0,0003358	0,09629	9,507e-16
4	0,2497153	0,0319962	0,0391041	0,0003754	0,08696	2,505e-14
5	0,3937821	0,0313725	0,0372162	0,0003891	0,19930	< 2,2e-16
6	0,4356194	0,0316358	0,0364886	0,0004076	0,23070	< 2,2e-16

Współrzędne prób wyznaczone za pomocą PCoA przedstawione w Tabeli 6.4 poddano analizie wariancji według modelu dla doświadczenia trójczynnikowego. Tabela 6.6 przedstawia wyniki tej analizy wariancji. W kolumnach PC1, PC2, PC3 i PC4 znajdują się wartości P dla statystyki F oraz wartości P otrzymane w wielowymiarowej analizie wariancji. Drukiem pogrubionym zostały zaznaczone wartości P < 0,05.

Tabela 6.6. Analiza wariancji dla danych NGS

Źródło zmienności	Wartości P z ANOVA				Wartość P z MANOVA
	PC1	PC2	PC3	PC4	
Genotyp (G)	0,7521	0,1904	**0,0142**	0,9073	0,0722
Dzień (D)	0,3973	0,1950	**0,0487**	**2,05e-05**	**8,809e-06**
Temperatura (T)	**1,68e-05**	0,3774	**0,0132**	0,0882	**1,699e-08**
Interakcja G×D	0,4998	0,7668	0,9446	0,9586	0,7053
Interakcja G×T	0,0646	0,9321	0,9340	**0,0170**	**0,0187**
Interakcja D×T	0,7270	**0,0463**	0,1364	**0,0037**	**3,481e-05**

Położenie prób w przestrzeni współrzędnych PCoA ilustruje Ryc. 6.6. Ponieważ w wynikach dla wszystkich współrzędnych wystąpiły wartości P mniejsze od ustalonego progu istotności 0,05, wykonane zostały wykresy dla PC1 w parze z: PC2, PC3 oraz PC4.

Poniżej podano interpretację wyników analizy wariancji i położenia prób w układzie współrzędnych głównych w kontekście czynników doświadczalnych zastosowanych w doświadczeniu i ich interakcji.

Ryc. 6.6.A: punkty odpowiadające optymalnej temperaturze (kolor czarny i szary) mają na ogół mniejsze wartości współrzędnych PC1 niż punkty odpowiadające wysokiej temperaturze (kolor

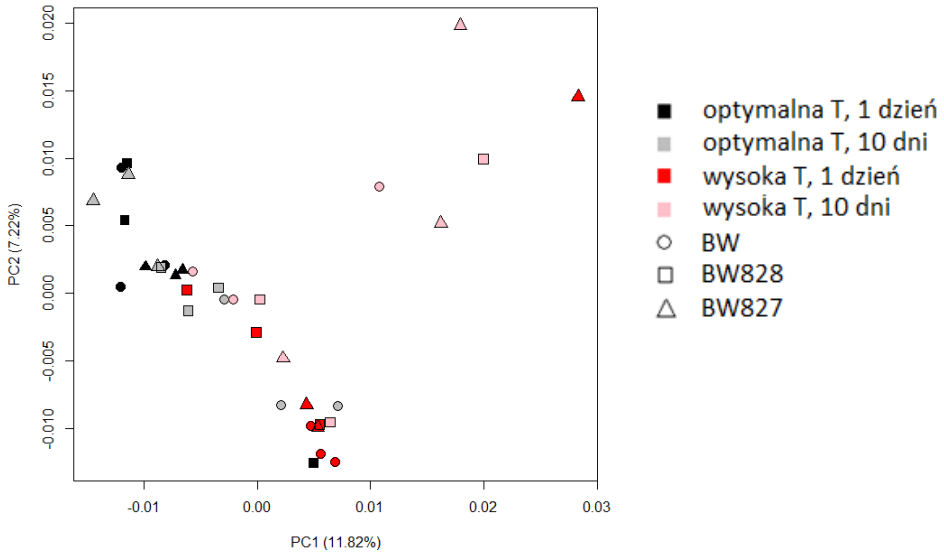
czerwony i różowy), co jest zgodne z istotnością efektu temperatury pokazaną w Tab. 6.6. Na osi PC2, dla grupy wyznaczonej przez wysoką temperaturę, punkty odpowiadające pierwszemu dniowi krzewienia (czerwone) mają na ogół niższe wartości niż te odpowiadające dziesiątemu dniowi (różowe), natomiast dla grupy wyznaczonej przez optymalną temperaturę punkty odpowiadające pierwszemu dniowi krzewienia (czarne) mają na ogół wyższe wartości niż te odpowiadające dziesiątemu dniowi krzewienia (szare); jest to zgodne z istotnością interakcji D×T dla PC2 pokazaną w Tab. 6.6.

Ryc. 6.6.B: układ punktów względem osi PC1 jest taki sam jak na Ryc. 6.6.A. Analizując układ punktów względem osi PC3 dochodzimy do wniosku, że punkty odpowiadające genotypowi BW mają współrzędne przeważnie ujemne, a punkty odpowiadające BW828 – przeważnie dodatnie. Punkty reprezentujące BW827 znajdują się po obu stronach osi PC3. Z tego wynika, że istotność efektu genotypu w Tab. 6.6 dla PC3 jest spowodowana głównie różnicą transkryptomów dla genotypów BW i BW828. Ponadto, punkty reprezentujące HT są w większości w górnej części wykresu, a reprezentujące OT – w dolnej. Jest to zgodne z istotnością efektu temperatury dla PC3. Istotny efekt dnia nie jest wyraźnie widoczny na Ryc. 6.6.B.

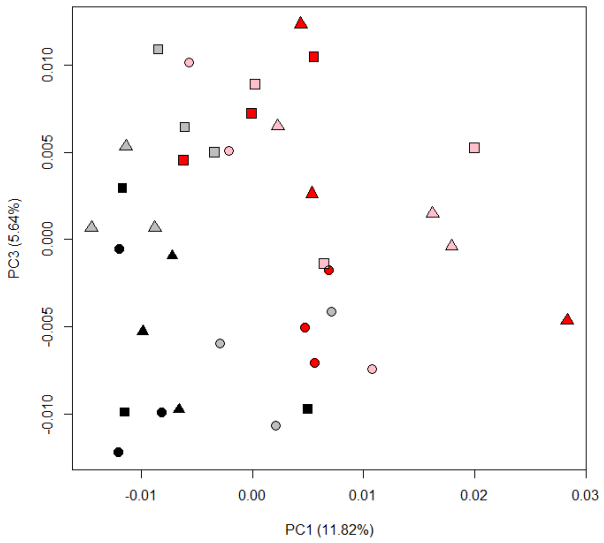
Ryc. 6.6.C: układ punktów względem osi PC1 taki jak na Ryc. 6.6.A. Patrząc na PC4 widzimy, że punkty odpowiadające pierwszemu dniowi krzewienia (punkty czerwone i czarne) mają na ogół mniejsze wartości niż punkty odpowiadające dziesiątemu dniowi (punkty jasne: różowe i szare). Jednakże, jak wskazuje istotność interakcji D×T w Tab. 6.6, podział wg dni wzdłuż osi PC4 nie jest jednakowy dla grup wyznaczonych przez temperaturę: w obrębie prób traktowanych wysoką temperaturą (HT) punkty różowe (10d) mają większe wartości niż czerwone (1d), natomiast w obrębie prób uzyskanych w optymalnej temperaturze (OT) taki podział jest mniej widoczny. Ponadto można zauważyć, że różnica pomiędzy punktami reprezentującymi HT i OT jest najbardziej wyraźna dla BW828, co tłumaczy istotność interakcji G×T dla PC4.

Wyniki MANOVA były zgodne z analizą poszczególnych współrzędnych i pokazały, że głównymi źródłami zmienności były dzień i temperatura oraz interakcja tych czynników.

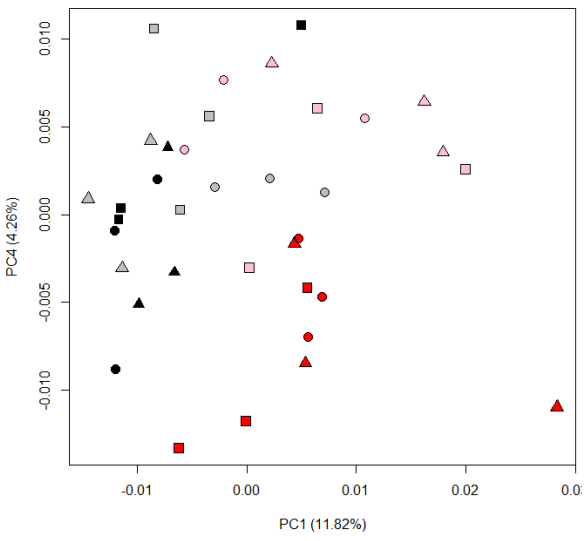
A.



B.



C.



Ryc. 6.6. Położenie prób w układzie współrzędnych głównych PC1, PC2, PC3 i PC4 otrzymanych na podstawie analizy PCoA danych NGS.

6.1.2. Analiza odległości wykorzystująca skafoldy

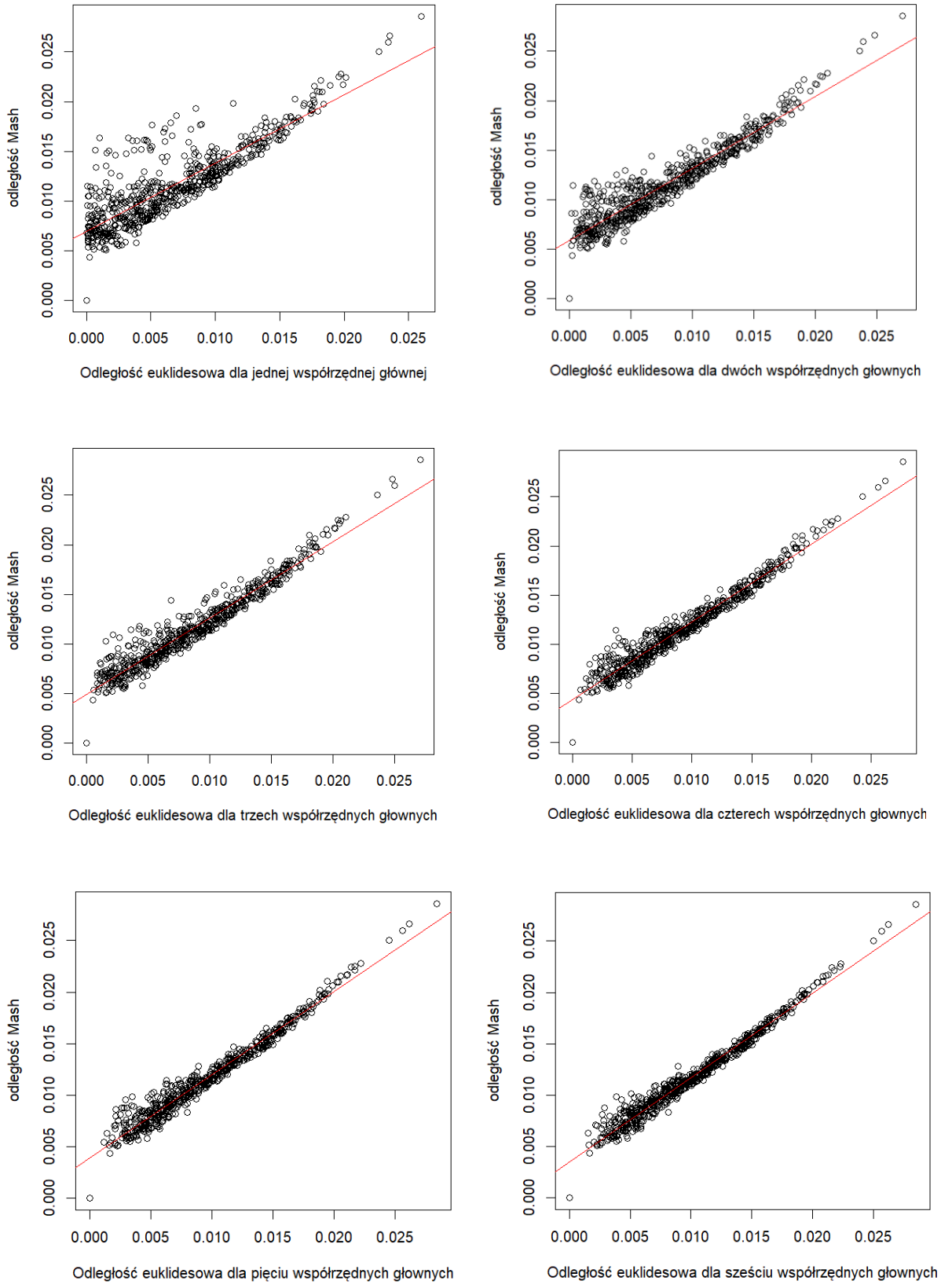
W podobny sposób analiza PCoA została wykonana dla zbiorów skafoldów, a następnie ANOVA oraz MANOVA. Procent zmienności wyjaśnianej przez pierwsze pięć współrzędnych głównych wynosił:

46,65%, 11,78%, 7,45%, 5,63% oraz 4,10%; całkowity dla tych 5 zmiennych: 75,61%. Wektory współrzędnych prób uzyskane z PCoA dla skafoldów zostały przedstawione w Tabeli 6.7.

Tabela 6.7. Wektory współrzędnych prób uzyskane z PCoA, pierwsze pięć współrzędnych głównych.

	PC1	PC2	PC3	PC4	PC5
BWOT1dR1	-0,0049	-0,0008	-0,0028	-0,0011	0,0005
BWOT1dR2	-0,0126	-0,0015	-0,0018	-0,0011	0,0056
BWOT1dR3	-0,0055	-0,0007	-0,0041	-0,0008	0,0008
BWOT10dR1	0,0005	-0,0003	0,0017	-0,0003	-0,0008
BWOT10dR2	0,0033	-0,0012	0,0018	0,0002	-0,0015
BWOT10dR3	0,0036	-0,0011	0,0013	0,0002	0,0001
BWHT1dR1	0,0073	-0,0036	0,0002	-0,0008	-0,0007
BWHT1dR2	0,0049	-0,0018	0,0005	-0,0017	-0,0012
BWHT1dR3	0,0035	-0,0015	-0,0010	-0,0018	-0,0012
BWHT10dR1	0,0003	0,0008	0,0007	0,0014	0,0010
BWHT10dR2	0,0007	0,0022	-0,0002	0,0007	-0,0011
BWHT10dR3	-0,0004	0,0021	0,0015	0,0019	-0,0001
BW827OT1dR1	-0,0101	-0,0017	-0,0028	0,0017	-0,0012
BW827OT1dR2	-0,0001	-0,0023	-0,0026	0,0007	0,0005
BW827OT1dR3	-0,0063	-0,0010	-0,0038	0,0005	-0,0018
BW827OT10dR1	-0,0048	0,0004	0,0052	0,0001	0,0013
BW827OT10dR2	-0,0100	0,0014	0,0050	-0,0013	-0,0025
BW827OT10dR3	-0,0035	0,0010	0,0028	-0,0011	-0,0014
BW827HT1dR1	0,0134	0,0062	-0,0023	-0,0065	-0,0009
BW827HT1dR2	0,0048	-0,0020	-0,0001	-0,0014	0,0004
BW827HT1dR3	0,0054	-0,0027	0,0001	-0,0006	0,0014
BW827HT10dR1	0,0049	-0,0011	0,0000	0,0024	0,0004
BW827HT10dR2	-0,0012	0,0132	-0,0023	0,0016	0,0003
BW827HT10dR3	0,0051	0,0011	0,0014	0,0031	0,0006
BW828OT1dR1	-0,0062	0,0005	-0,0020	0,0009	0,0003
BW828OT1dR2	-0,0093	-0,0003	-0,0028	-0,0008	-0,0042
BW828OT1dR3	0,0039	-0,0029	-0,0023	0,0006	-0,0016
BW828OT10dR1	-0,0005	0,0001	0,0032	0,0005	0,0000
BW828OT10dR2	-0,0028	0,0001	0,0030	0,0010	-0,0004
BW828OT10dR3	-0,0043	0,0004	0,0034	-0,0002	0,0002
BW828HT1dR1	0,0058	-0,0035	-0,0003	-0,0005	0,0008
BW828HT1dR2	0,0027	-0,0012	-0,0002	-0,0016	0,0031
BW828HT1dR3	-0,0045	0,0005	0,0016	-0,0050	0,0017
BW828HT10dR1	0,0037	-0,0004	-0,0005	0,0019	0,0014
BW828HT10dR2	0,0055	-0,0015	-0,0009	0,0035	-0,0023
BW828HT10dR3	0,0075	0,0029	-0,0008	0,0037	0,0025

Tak jak dla surowych danych, dla skafoldów wykonane zostały wykresy Sheparda dla wymiarów od jednego do sześciu (Ryc. 6.7).



Ryc. 6.7. Wykresy Sheparda dla skafoldów, kolejno dla jednej, dwóch, trzech, czterech, pięciu i sześciu pierwszych współrzędnych.

Została również przeprowadzona regresja liniowa, której wyniki przedstawione zostały w tabeli poniżej (Tabela 6.8).

Tabela 6.8. Wyniki regresji liniowej dla odległości Mash i odległości euklidesowej.

	Współczynniki regresji i ich błędy standardowe (regresja liniowa; funkcja postaci: $y = ax + b$)					
Liczba wymiarów	*a*	błąd standardowy	*b*	błąd standardowy	Współczynnik determinacji	Wartość P dla statystyki F
1	0,6857542	0,0157909	0,0069496	0,0001316	0,7498	2,2e-16
2	7,268e-01	9,916e-03	5,874e-03	9,247e-05	0,8952	2,2e-16
3	7,666e-01	8,263e-03	4,967e-03	8,183e-05	0,9319	2,2e-16
4	7,891e-01	6,952e-03	4,364e-03	7,172e-05	0,9534	< 2,2e-16
5	8,027e-01	6,237e-03	3,953e-03	6,617e-05	0,9634	< 2,2e-16
6	8,190e-01	5,480e-03	3,531e-03	5,957e-05	0,9726	2,2e-16

Tabela 6.9 przedstawia wyniki analizy wariancji. Tak jak w Tabeli 6.6, tu również w kolumnach PC1, PC2, PC3 i PC4 znajdują się wartości P otrzymane w wyniku jednowymiarowej analizy wariancji, a w kolumnie MANOVA znajdują się wartości P dla wielowymiarowej analizy wariancji, wszystkie zaokrąglone do czterech miejsc po przecinku. Pogrubionym drukiem zostały zaznaczone wartości P < 0,05.

Tabela 6.9. Analiza wariancji dla skafoldów

Źródło zmienności	Wartości P z ANOVA				Wartość P z MANOVA
	PC1	PC2	PC3	PC4	
genotyp (G)	0,9794	0,2891	0,7600	0,5929	0,4717
dzień (D)	0,5550	**0,0238**	**5,2263e-09**	**0,0001**	**4,408e-12**
temperatura (T)	**1,3123e-05**	0,2501	0,4274	0,9548	**6,723e-05**
Interakcja G×D	0,2278	0,7138	0,1281	0,6033	0,4961
Interakcja G×T	0,2357	0,3987	0,0522	0,8020	0,1521
Interakcja D×T	0,0657	0,3098	**8,1162e-09**	**4,0577e-05**	**2,232e-09**

Tak jak w przypadku danych NGS, położenie prób w przestrzeni współrzędnych PCoA zostało zilustrowane po uprzednim sprawdzeniu, dla których współrzędnych pojawiają się wartości P mniejsze od ustalonego progu (Ryc. 6.8).

ANOVA i MANOVA wskazały, podobnie jak dla danych NGS, istotność wpływu temperatury, dnia obserwacji oraz interakcji tych czynników na sekwencje transkryptów.

Ryc.6.8.A: patrząc na PC1 obserwujemy podobny układ punktów jak w przypadku PC1 dla surowych danych. Tutaj również punkty odpowiadające optymalnej temperaturze (czarne i szare) mają w ogólności niższe wartości niż punkty odpowiadające wysokiej temperaturze (czerwone i różowe).

Punkty odpowiadające pierwszemu dniowi krzewienia (czarne i czerwone) mają ogólnie niższe wartości PC2 niż te odpowiadające dziesiątemu dniowi (szare i różowe).

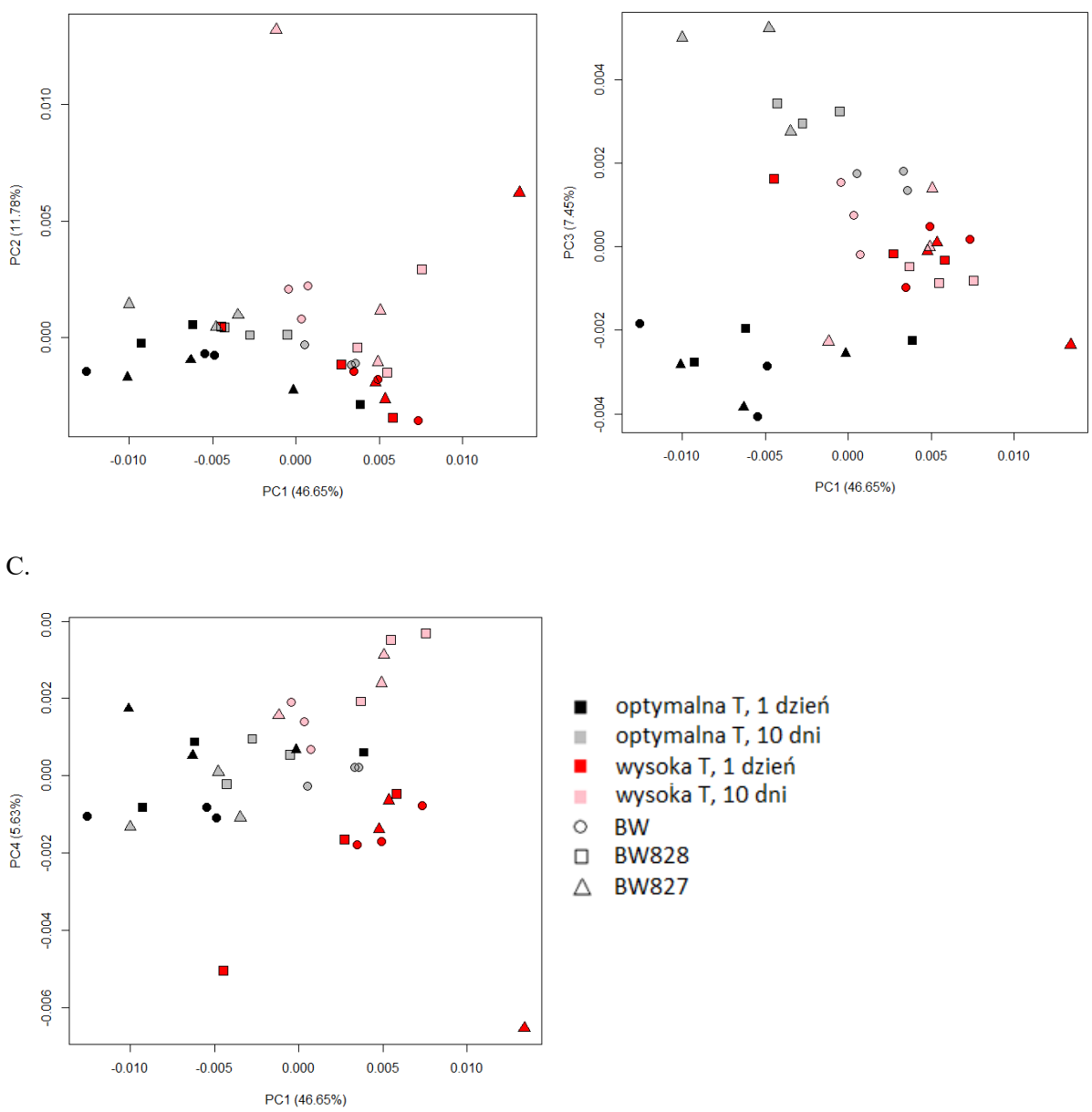
Ryc.6.8.B: układ punktów względem osi PC1 jest taki jak na Ryc. 6.8.A. Patrząc na PC3 widzimy, że wartości odpowiadające pierwszemu dniowi krzewienia (czarne i czerwone) mają niższe wartości niż te odpowiadające dziesiątemu dniowi (szare i różowe), jednak rozpatrując osobno grupy wyznaczone przez różne temperatury widzimy, że ze względu na czynnik dnia punkty odpowiadające zbiorom danych rozłożone są inaczej: dla optymalnej temperatury (czarne i szare) punkty odpowiadające pierwszemu dniowi mają niższe wartości niż punkty odpowiadające dziesiątemu dniowi, natomiast dla wysokiej temperatury (czerwone i różowe) przedziały wartości dla obu punktów czasowych są podobne.

Ryc.6.8.C: układ punktów względem osi PC1 jest taki jak na Ryc. 6.8.A. Patrząc na PC4 możemy zaobserwować, że punkty odpowiadające pierwszemu dniowi krzewienia mają na ogół niższe wartości niż te odpowiadające dziesiątemu dniowi. Jednocześnie jeśli rozważymy rozkłady punktów wewnątrz grup temperaturowych, widzimy że dla wysokiej temperatury (czerwone i różowe) punkty odpowiadające pierwszemu dniowi mają niższe wartości niż te odpowiadające dziesiątemu dniowi, natomiast dla temperatury optymalnej przedziały wartości dla obu punktów czasowych są podobne.

Zauważmy, że obraz interakcji D×T jest inny dla wizualizacji otrzymanych dla współrzędnych PC3 i PC4. Na Ryc. 6.8.B, punkty odpowiadające OT, lecz różnym momentom obserwacji, są oddalone, natomiast punkty odpowiadające HT są położone blisko. Na Ryc. 6.8.C sytuacja jest odwrotna. Tak więc wyniki analizy dla skafoldów pokazują, że istnieje wymiar (PC3) pokazujący większe zróżnicowanie dni w ramach OT niż w ramach HT, oraz wymiar (PC4) pokazujący odwrotną zależność.

A.

B.



C.

Ryc. 6.8. Położenie prób w układzie współrzędnych głównych PC1, PC2, PC3 i PC4 otrzymanych na podstawie analizy PCoA dla skafoldów.

W celu sprawdzenia, czy etap składania, na którym były zbiory danych, wpływał na ich rozmieszczenie po analizie współrzędnych głównych, przeprowadzono jeszcze jedną analizę wariancji, tym razem uwzględniającą czynnik „przed/po składaniu" jako jeden z czynników eksperymentalnych. Wyniki obliczeń zamieszczone są w tabeli poniżej (Tabela 6.10).

Dla wszystkich czterech współrzędnych pojawiły się wartości P mniejsze od założonego progu. Na rycinie 6.9 przedstawiono położenie zbiorów danych względem siebie dla PC1 i PC2 oraz dla PC3 i PC4.

Tabela 6.10. Analiza wariancji dla pierwszych czterech współrzędnych głównych przy uwzględnieniu etapu składania jako jednego ze źródeł zmienności.

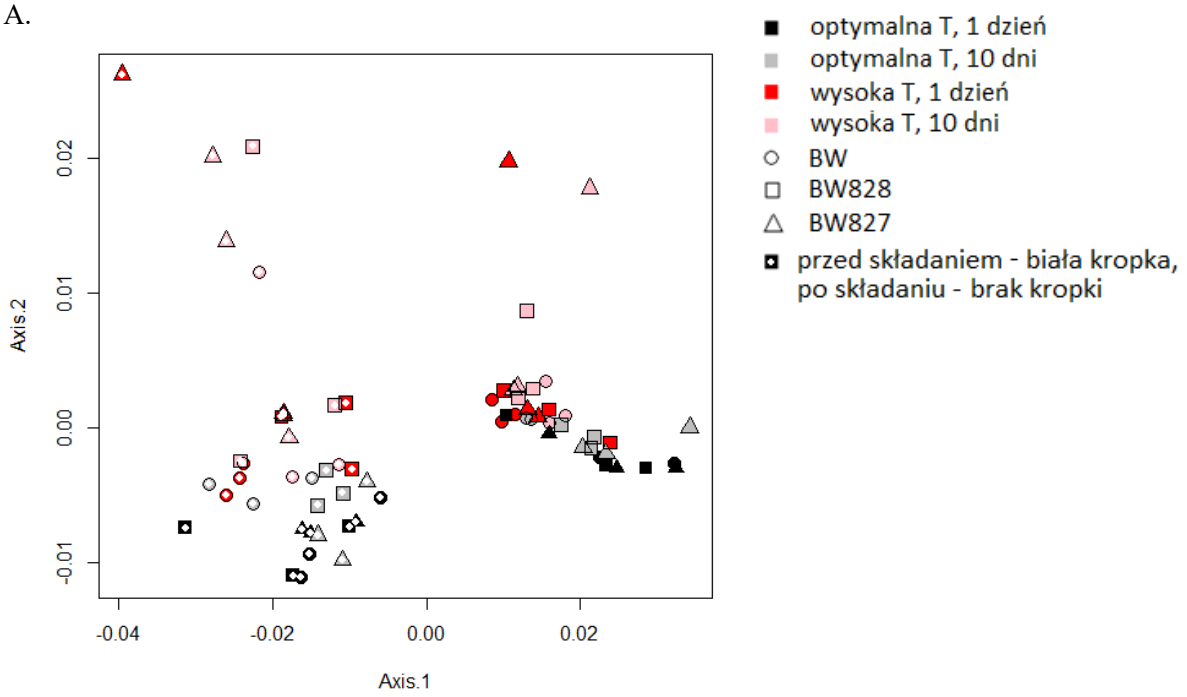
Źródło zmienności	ANOVA				MANOVA
	PC1	PC2	PC3	PC4	Wartość P
Przed/Po Składaniu (P)	**7,2228e-31**	**0,0493**	0,2070	0,2753	**2,2e-16**
Genotyp (G)	0,4708	0,0589	0,0953	**0,0011**	**0,0015**
Dzień (D)	0,8711	0,0533	0,4220	**5,1352e-06**	**7,756e-06**
Temperatura (T)	**9,9960e-06**	**2,1269e-07**	**6,4479e-05**	0,7966	**2,893e-13**
Interakcja P×G	0,4772	0,7409	0,5926	**0,0219**	0,1194
Interakcja P×D	0,5861	0,4318	0,9495	**0,0018**	**0,0127**
Interakcja P×T	0,3499	**0,0327**	**0,0044**	0,4436	**1,470e-06**
Interakcja G×D	0,5163	0,6347	0,2373	0,7557	0,3990
Interakcja G×T	**0,0259**	**0,0167**	0,8172	0,0934	**0,0029**
Interakcja D×T	0,2610	0,7427	**0,0001**	0,4470	**0,0005**
Interakcja P×G×D	0,9536	0,8711	0,5252	0,9629	0,7924
Interakcja P×G×T	0,5934	0,7461	0,8689	0,3024	0,6748
Interakcja P×D×T	0,4358	0,6962	**0,0472**	0,7184	0,1545
Interakcja G×D×T	**0,0010**	0,8783	0,0653	**0,0326**	**0,0168**

Ryc. 6.9.A: patrząc na PC1 widzimy, że punkty odpowiadające zbiorom danych przed składaniem (punkty z białą kropką w środku) mają niższe wartości niż te odpowiadające skafoldom (punkty pełne). Ze względu na temperaturę, ogólnie punkty odpowiadające optymalnej temperaturze (czarne i szare) mają nieco wyższe wartości niż te odpowiadające wysokiej temperaturze (czerwone i różowe). Różni się ta sytuacja między genotypami, np. dla BW827 jest to bardziej widoczne niż dla BW828, co może potwierdzać interakcję czynników genotyp × temperatura dla PC1. Patrząc oddzielnie na każdy zbiór punktów wyznaczony przez kolor (czerwone, czarne, różowe, szare) rozważymy podzbiory wyznaczone przez dzień i temperaturę. Wówczas można zauważyć, że w tych podzbiorach punkty odpowiadające różnym genotypom różnie są względem siebie rozmieszczone. Na przykład dla wysokiej temperatury pierwszego dnia (punkty czerwone) Bowman ma na ogół niższe wartości niż BW827, czego nie można stwierdzić o tych samych genotypach dla wysokiej temperatury dziesiątego dnia (punkty różowe). Punkty odpowiadające wysokiej temperaturze mają na ogół wyższe wartości PC2 niż te odpowiadające optymalnej temperaturze. Względem tej osi nieco mniej widać różnice położenia punktów względem etapu składania, jednak nadal można zauważyć, że punkty odpowiadające skafoldom mają na ogół niższe wartości niż te odpowiadające surowym danym NGS. Analiza wariancji stwierdziła jeszcze istotność dwóch interakcji dla PC2: etap składania × temperatura oraz genotyp × temperatura. Patrząc na różnice między punktami odpowiadającymi wysokiej i optymalnej temperaturze widzimy, że dla zbiorów danych przed składaniem są one większe niż dla tych po składaniu. Patrząc na genotyp zauważamy, że różnice między temperaturami są ogólnie większe dla BW827 (trójkąty) i dla BW828 (kwadraty) niż dla Bowmana (kółka).

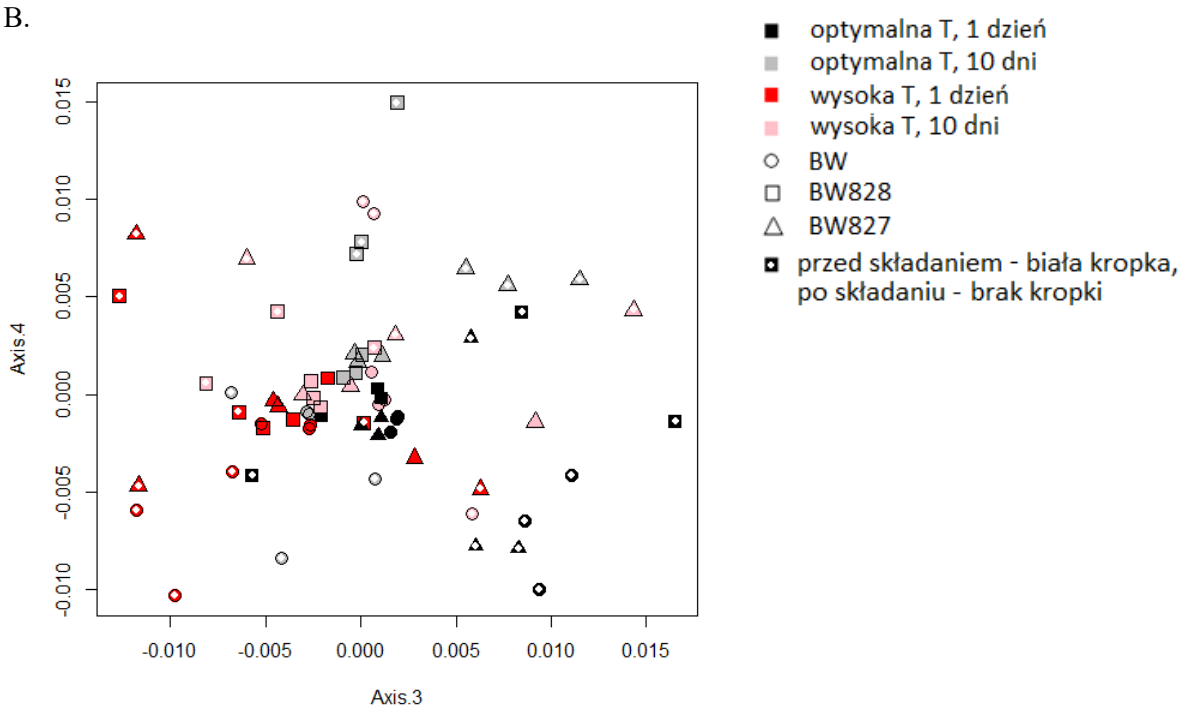
Ryc. 6.9.B: punkty odpowiadające optymalnej temperaturze mają ogólnie wyższe wartości PC3 niż te odpowiadające wysokiej temperaturze. Patrząc pod kątem etapu składania, dla punktów odpowiadających surowym danym (punkty z kropką) bardziej widać różnicę ze względu na temperaturę niż dla skafoldów (punkty pełne). Biorąc pod uwagę czynnik dzień, dla dnia pierwszego (punkty czerwone i czarne) bardziej widać zróżnicowanie rozłożenia punktów względem temperatury niż dla

dnia dziesiątego (punkty różowe i szare). Patrząc na punkty odpowiadające różnym temperaturom, dla pierwszego dnia można zauważyć większą różnicę niż dla dziesiątego dnia, przy czym dla surowych danych jest to większa różnica niż dla skafoldów. W przypadku PC4 punkty odpowiadające pierwszemu dniowi (różowe i szare) mają na ogół wyższe wartości niż te odpowiadające dziesiątemu dniowi (czerwone i czarne). Biorąc pod uwagę zarówno dzień jak i etap składania, punkty odpowiadające etapowi przed składaniem (punkty z białą kropką) bardziej się różnią wartościami niż po (punkty pełne). Patrząc na genotyp, punkty odpowiadające Bowmanowi (BW, kółka) mają w ogólności niższe wartości niż pozostałe dwa genotypy (BW828 – kwadraty, BW827 - trójkąty). Jednocześnie dla punktów odpowiadających zbiorom danych przed składaniem jest to zależność bardziej widoczna niż dla punktów odpowiadających skafoldom. Żeby dojrzeć interakcję trzech czynników: genotypu, dnia i temperatury warto przyjrzeć się oddzielnie punktom w każdym kolorze – odpowiadają one konkretnemu dniowi i temperaturze – i przeanalizować ich położenie z uwzględnieniem genotypu (kształty punktów). Istotnie dostrzec można różnice w położeniu względem siebie punktów odpowiadających poszczególnym genotypom w zależności od tego, jaki dzień i temperaturę analizujemy, na przykład dla pierwszego dnia i wysokiej temperatury (punkty czerwone) genotyp BW przyjmuje na ogół niższe wartości niż pozostałe genotypy, natomiast dla drugiego dnia i wysokiej temperatury (różowe punkty) nie zachodzi taka zależność.

A.



B.



Ryc. 6.9. Położenie prób w układzie współrzędnych głównych PC1, PC2, PC3 i PC4 otrzymanych na podstawie analizy PCoA dla surowych danych NGS i skafoldów.

Odległości Mash policzone dla par zbiorów: surowe dane NGS - złożone z nich skafoldy, przedstawia Tabela 6.11. Wyniki znalazły się w przedziale między 0,0346 (BW828HT10dR1) a 0,0544 (BW828OT1dR2).

Tabela 6.11. Odległość Mash między surowymi danymi i skafoldami.

Próba	Odległość	Próba	Odległość	Próba	Odległość
BWOT1dR1	0,0486	BW827OT1dR1	0,0477	BW828OT1dR1	0,0429
BWOT1dR2	0,0452	BW827OT1dR2	0,0420	BW828OT1dR2	0,0544
BWOT1dR3	0,0475	BW827OT1dR3	0,0477	BW828OT1dR3	0,0475
BWOT10dR1	0,0410	BW827OT10dR1	0,0406	BW828OT10dR1	0,0411
BWOT10dR2	0,0482	BW827OT10dR2	0,0519	BW828OT10dR2	0,0438
BWOT10dR3	0,0452	BW827OT10dR3	0,0424	BW828OT10dR3	0,0439
BWHT1dR1	0,0399	BW827HT1dR1	0,0465	BW828HT1dR1	0,0384
BWHT1dR2	0,0427	BW827HT1dR2	0,0427	BW828HT1dR2	0,0355
BWHT1dR3	0,0430	BW827HT1dR3	0,0411	BW828HT1dR3	0,0429
BWHT10dR1	0,0373	BW827HT10dR1	0,0391	BW828HT10dR1	0,0346
BWHT10dR2	0,0449	BW827HT10dR2	0,0482	BW828HT10dR2	0,0435
BWHT10dR3	0,0436	BW827HT10dR3	0,0447	BW828HT10dR3	0,0435

W celu sprawdzenia, czy któreś z czynników eksperymentalnych wpłynęły istotnie na odległości między surowymi danymi a skafoldami, przeprowadzona została analiza wariancji (Tab. 6.12). Przyjęto $P < 0,05$. Wyniki wskazały istotność czynnika temperatura. Średnia odległość w grupie HT wynosiła 0,042 (bł. std. 0,00087), w grupie OT 0,046 (bł. std. 0,00091). Tak więc składanie sekwencji bardziej oddaliło skafoldy od danych NGS dla prób pobranych w warunkach OT niż dla prób pobranych w warunkach HT.

Tabela 6.12. ANOVA dla odległości Mash dla par NGS-skafoldy.

Czynnik	Wartość P
Genotyp (G)	0,4885
Dzień (D)	0,4281
Temperatura (T)	**0,0060**
G×D	0,8604
G×T	0,4236
D×T	0,1788

6.2. Analiza podobieństwa transkryptomów w aspekcie funkcji molekularnych i procesów biologicznych

Dotąd rozważaliśmy odległości Mash między zbiorami danych w celu sprawdzenia, czy różnice między sekwencjami w sensie występujących albo niewystępujących ciągów nukleotydów mogą być spowodowane konkretnymi czynnikami eksperymentalnymi. W tej analizie nie były brane pod uwagę funkcje sekwencji, a jedynie ich budowa. W tym rozdziale zajmiemy się próbą połączenia adnotacji sekwencji z ich interpretacją w kontekście analizy wykorzystującej odległość Mash. Podejmiemy się tego zadania korzystając z Gene Ontology (GO), a dokładniej z uogólnienia GO: GOSlim GOA z Ensembl Plants, omówionej w rozdziale 4.6.

W celu przeprowadzenia dalszych badań, z bazy Ensembl Plants pobraliśmy transkryptom referencyjny *Hordeum vulgare* wraz ze wszystkimi izoformami (wersja z dnia 07.04.2021). Korzystając ze skryptu napisanego w języku R stworzyliśmy listę wszystkich terminów GOSlim występujących w adnotacji sekwencji referencyjnych. Łącznie lista zawierała 96 terminów. Za pomocą narzędzia blast (blast+/2.7.1; parametr -evalue 10e-50) zmapowaliśmy w sekwencjach referencyjnych sekwencje skafoldów zawarte w zbiorach danych eksperymentalnych, w celu otrzymania identyfikatorów GOSlim GOA dla tych skafoldów (wybierane były te o najwyższym procencie zmapowania). Następnie podzieliliśmy każdy zbiór danych na zbiory zawierające tylko te skafoldy, które zostały przypisane konkretnym terminom w GOSlim GOA. Dla 4 terminów wystąpiły zbiory puste. Ostatecznie każdy zbiór skafoldów podzieliliśmy na 92 podzbiory, każdy zawierający tylko te sekwencje, które zostały przypisane konkretnemu terminowi.

Kolejnym krokiem było obliczenie odległości Mash dla tych podzbiorów. Podobnie jak w poprzedniej analizie, na macierzach została przeprowadzona PCoA oraz analiza wariancji w celu sprawdzenia, czy któryś z czynników eksperymentalnych miał statystycznie istotny wpływ na rozmieszczenie punktów względem osi uzyskanych przez PCoA: PC1, PC2, PC3 lub PC4. Ze względu na wielokrotne porównania przy testowaniu hipotez w ANOVA zastosowano poprawkę Bonferroniego. Wynikiem są cztery tabele (Suplement, Tabela S3, S4, S5, S6) przedstawiające wartości P dla pierwszych czterech osi dla każdego z analizowanych terminów GOSlim oraz tabela przedstawiająca procent zmienności objaśnianej przez te osie (Suplement, Tabela S2).

W Tabeli S2 zawarliśmy informacje na temat procentu zmienności objaśnianej przez osie PC1, PC2, PC3 i PC4. Dla pierwszej współrzędnej wartości zawierały się w przedziale między 0,2029 (*thylakoid*) a 0,8971 (*motor activity*), średnia tych wartości wyniosła 0,4527, a mediana 0,4406. Dla drugiej współrzędnej minimalna wartość wyniosła 0,0474 (*motor activity*), maksimum 0,2859 (*circadian rhythm*), średnia 0,1011, a mediana 0,0900. Dla trzeciej współrzędnej statystyki te wynosiły odpowiednio 0,0367 (*motor activity*), 0,1688 (*signaling receptor binding*), 0,0712 oraz 0,0666, a dla czwartej współrzędnej 0,0295 (*DNA metabolic process*), 0,1179 (*signaling receptor binding*), 0,0553 oraz 0,0533.

W ramach ontologii Molecular Function, największy procent zmienności transkryptomów opisany był przez PC1 dla terminów *nuclease activity* i *chromatin binding*. Należy wnioskować, iż dla tych terminów zmienność sekwencji może być najlepiej opisana w jednym wymiarze. Podobnie, dla ontologii Cellular Component największy procent zmienności opisywanej przez PC1 dotyczył terminów *nuclear envelope* i *nucleoplasm*, a dla ontologii biological process - terminów *motor activity* i *DNA metabolic process*.

Spójrzmy dalej na tabele z wartościami P otrzymanymi w ANOVA dla kolejno PC1, PC2, PC3 oraz PC4.

Patrząc na tabelę dla pierwszej współrzędnej głównej widzimy, że dla żadnego terminu genotyp nie był czynnikiem wpływającym istotnie na zmienność punktów odpowiadających zbiorom danych. Dzień był istotny jedynie dla *secondary metabolic process*, natomiast dla 44 terminów (*binding, biological process, biosynthetic process, carbohydrate binding, carbohydrate metabolic process, catabolic process, catalytic activity, cell wall, cellular component, cellular process, chloroplast, circadian rhythm, DNA binding, DNA-binding transcription factor activity, endoplasmic reticulum, endosome, enzyme regulator activity, external encapsulating structure, extracellular region, generation of precursor metabolites and energy, hydrolase activity, lipid metabolic process, membrane, metabolic process, mitochondrion, molecular function, photosynthesis, plasma membrane, plastid, protein metabolic process, regulation of molecular function, response to biotic stimulus, response to chemical, response to stress, secondary metabolic process, signaling receptor binding, structural molecule activity, thylakoid, transcription regulator activity, translation, translation factor activity RNA binding, translation regulator activity, transporter activity* oraz *vacuole*) istotny był czynnik temperatury. Z interakcji jedynie dla *extracellular region* interakcja dnia z temperaturą okazała się istotna.

Dla drugiej współrzędnej głównej genotyp był istotny tylko dla *endoplasmic reticulum* oraz *Golgi apparatus*. Dzień wpływał istotnie w 49 przypadkach (*binding, biological process, carbohydrate binding, carbohydrate metabolic process, catabolic process, catalytic activity, cell communication, cell wall, cellular component, cellular component organization, cellular homeostasis, cellular process, cellular protein modification process, DNA-binding transcription factor activity, external encapsulating structure, extracellular region, flower development, hydrolase activity, kinase activity, lipid metabolic process, membrane, metabolic process, molecular function, multicellular organism development, nuclease activity, nucleic acid binding, nucleobase-containing compound metabolic process, nucleoplasm, nucleotide binding, nucleus, plasma membrane, plastid, pollination, post-embryonic development, protein binding, protein metabolic process, reproduction, response to abiotic stimulus, response to chemical, response to endogenous stimulus, response to light stimulus, response to stress, secondary metabolic process, signal transduction, signaling receptor activity, transcription regulator activity, transferase activity, transport oraz transporter activity*). Dla temperatury nie wykazano istotności, jednak istotność interakcji temperatury z czasem otrzymano w 34 przypadkach (*binding, biological process, biosynthetic process, carbohydrate binding, catabolic process, catalytic activity, cell cycle, cell differentiation, cell wall, cellular component, cellular component organization, cellular process, cytoskeleton, DNA binding, DNA metabolic process, enzyme regulator activity, external encapsulating structure, flower development, hydrolase activity, intracellular, metabolic process, molecular function, nuclease activity, nucleic acid binding, nucleobase-containing compound metabolic process, nucleus, protein binding, protein metabolic process, regulation of molecular function, response to abiotic stimulus, response to stress, transferase activity, transport* oraz *transporter activity*). Pozostałe rozważane interakcje okazały się nie wpływać na rozmieszczenie punktów względem PC2.

Dla trzeciej współrzędnej głównej genotyp był istotny tylko dla dwóch terminów: *secondary metabolic process* oraz *signaling receptor activity*. Dzień był istotnym czynnikiem w 18 przypadkach (*binding, biosynthetic process, carbohydrate binding, catalytic activity, cellular process, cellular protein modification process, DNA binding, DNA-binding transcription factor activity, extracellular region, generation of precursor metabolites and energy, hydrolase activity, metabolic process, mitochondrion, molecular function, nucleobase-containing compound metabolic process, nucleus, protein binding* oraz

thylakoid). Podobnie jak dla drugiej osi nie wykazano istotności temperatury, jednak przy interakcji temperatury z czasem otrzymano istotność w 29 przypadkach (*DNA-binding transcription factor activity, cellular protein modification process, membrane, extracellular region, kinase activity, multicellular organism development, catalytic activity, external encapsulating structure, molecular function, anatomical structure development, nucleic acid binding, cell wall, metabolic process, carbohydrate binding, mitochondrion, nucleotide binding, biological process, transcription regulator activity, embryo development, nucleobase-containing compound metabolic process, transferase activity, carbohydrate metabolic process, transport, protein binding, cellular component, reproduction, response to biotic stimulus, post-embryonic development* oraz *response to abiotic stimulus*). Ponadto dla interakcji genotypu z temperaturą wykazano istotność dla *DNA-binding transcription factor activity*.

Dla czwartej współrzędnej głównej genotyp był istotny dla terminu *reproduction*, temperatura dla *DNA binding*, a dzień okazał się nieistotny dla wszystkich terminów. Z badanych interakcji tylko interakcja dnia z temperaturą wykazała istotność – stało się tak dla 9 terminów (*cytosol, endosome, lipid metabolic process, response to chemical, response to endogenous stimulus, response to external stimulus, signal transduction, structural molecule activity* oraz *translation*).

Dysponując przedstawionymi tabelami można sprawdzić, dla których terminów warto przedstawić graficznie odległości między podzbiorami danych sekwencyjnych i przyjrzeć się zależności ich położenia od poziomów czynników eksperymentalnych.

Przykład 1: *secondary metabolic process*

Analiza została przeprowadzona podobnie jak w rozdziale 6.1, tym razem dla podzbiorów danych sekwencyjnych zawierających tylko sekwencje (skafoldy) związane z *secondary metabolic process*. Zatem, na początku policzone zostały odległości Mash dla każdej pary podzbiorów i tym sposobem uzyskana została macierz odległości. Następnie przeprowadzona została PCoA. ANOVA i MANOVA pozwoliły stwierdzić, czy na rozłożenie punktów reprezentujących zbiory danych w uzyskanym tym sposobem układzie względem współrzędnych PC1, PC2, PC3 i PC4 mogły mieć wpływ czynniki eksperymentalne: genotyp, temperatura, dzień lub interakcja tych czynników.

Tabela 6.12. Wyniki analizy wariancji dla podzbiorów *secondary metabolic process*.

Czynnik	Wartość P ANOVA				Wartość P MANOVA
	PC1 (27,19%)	PC2 (18,53%)	PC3 (9,78%)	PC4 (8,09%)	
Genotyp (G)	0,32323	0,83876	**6e-06**	0,06888	0,0008092
Dzień (D)	**5e-05**	**5e-05**	0,38616	0,43659	**2,012e-11**
Temperatura (T)	**9e-05**	0,00112	0,35904	0,01633	**4,385e-07**
G×D	0,41332	0,54475	0,33745	0,04287	0,0517117
G×T	0,25265	0,23898	0,86712	0,44982	0,0705189
D×T	0,00106	0,01026	0,01526	0,22339	**2,458e-09**

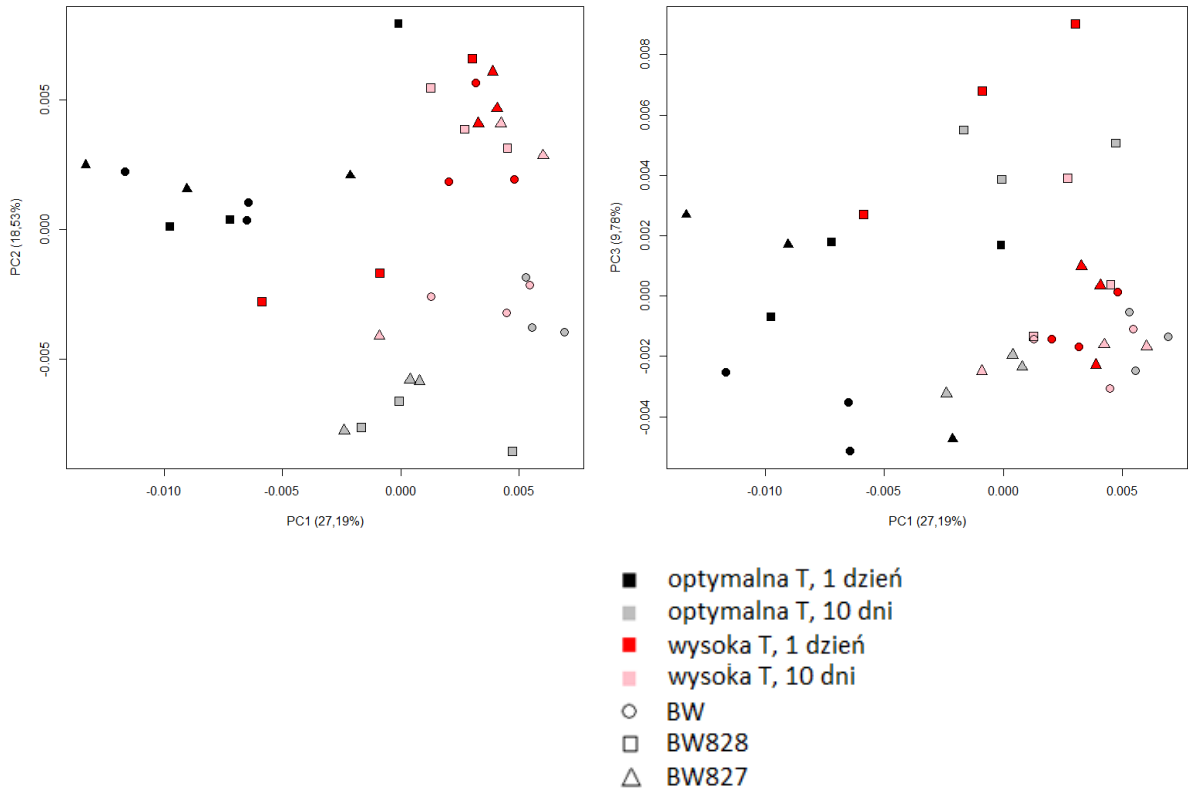
Wyniki wskazują na związek położenia punktów względem osi PC1 z dniem i temperaturą. Dla PC2 istotnym czynnikiem był dzień, natomiast dla PC3 genotyp. Dla PC4 nie wykazano istotności żadnego z czynników ani ich interakcji. Zauważmy, że dla interakcji czynników dzień i temperatura ANOVA nie wykazała istotności, a MANOVA – tak, natomiast dla czynnika genotyp nastąpiła sytuacja odwrotna: ANOVA wykazała istotność, a MANOVA nie. Poniżej (Ryc. 6.10) przedstawione są wykresy punktowe przedstawiające rozmieszczenie punktów reprezentujących podzbiory danych względem osi PC1, PC2 oraz PC3.

Ryc. 6.10.A: względem osi PC1 punkty związane z pierwszym dniem (różowe i szare) mają w ogólności niższe wartości niż te związane z dziesiątym dniem (czerwone i czarne), a punkty związane z wysoką temperaturą (czerwone i różowe) mają w ogólności wyższe wartości niż te związanie z optymalną (czarne i szare). Względem osi PC2 punkty związane z pierwszym dniem mają w ogólności wyższe wartości niż te związane z dziesiątym.

Ryc. 6.10.B: punkty odpowiadające genotypowi BW828 (kwadraty) mają w ogólności wyższe wartości PC3 niż punkty odpowiadające dwóm pozostałym genotypom (kółka, trójkąty).

A.

B.



Ryc. 6.10. Położenie prób w układzie współrzędnych głównych PC1, PC2 i PC3 otrzymanych na podstawie analizy PCoA dla podzbiorów skafoldów, dla *secondary metabolic process*.

Przykład 2: *extracellular region*

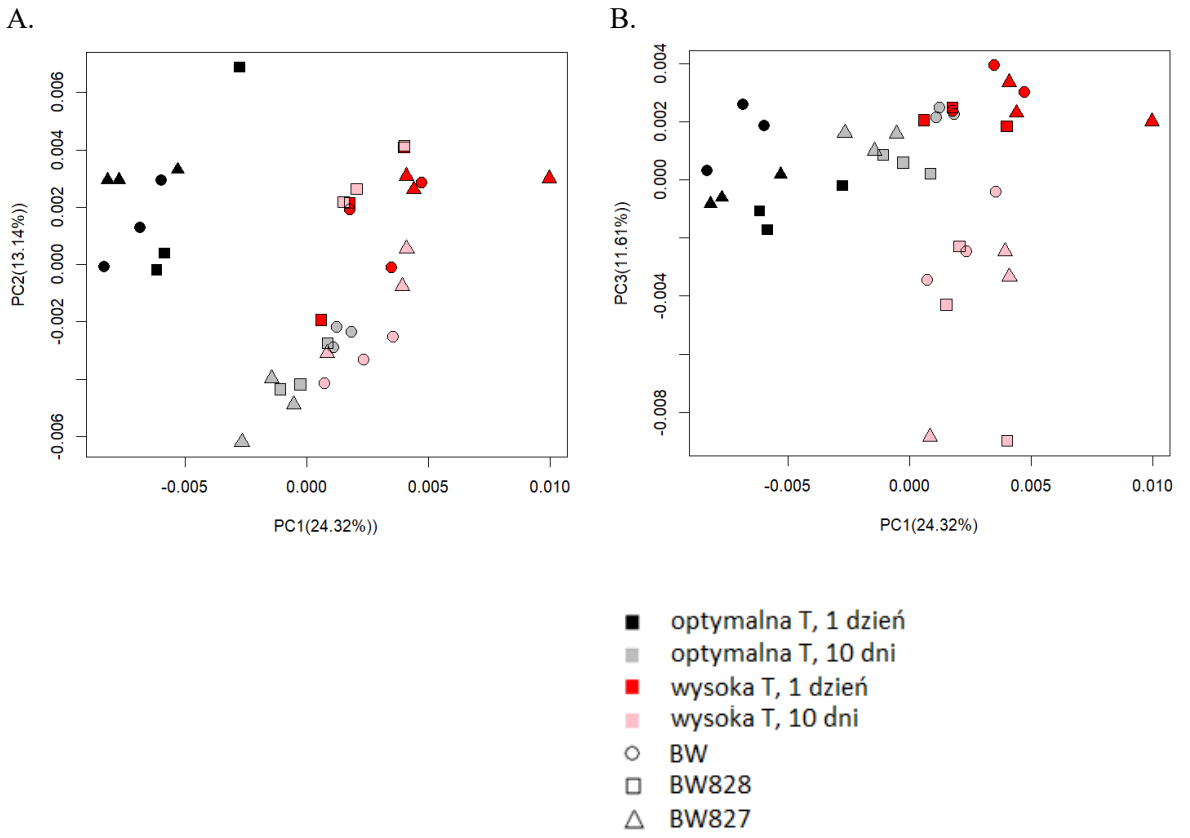
Analiza została przeprowadzona jak w pierwszym przykładzie. Wyniki przedstawione w tabeli poniżej (Tab. 6.13) wskazały na związek położenia punktów względem osi PC1 z temperaturą oraz z interakcją czynników temperatura i dzień. Dla PC2 istotnym czynnikiem był dzień, natomiast dla PC3 dzień oraz interakcja czynników dzień i temperatura. Dla PC4 nie wykazano istotności statystycznej żadnego z czynników ani ich interakcji. Wyniki MANOVA były zgodne z wynikami ANOVA. W związku z powyższym, stworzone zostały wykresy dla osi PC1, PC2 oraz PC3.

Tabela 6.13. Wyniki analizy wariancji dla podzbiorów *extracellular region*.

Czynnik	Wartość P ANOVA				Wartość P MANOVA
	PC1 (24,32%)	PC2 (13,14%)	PC3 (11,61%)	PC4 (6,31%)	
Genotyp (G)	0,9439	0,2084	0,0088	0,2377	0,077874
Dzień (D)	0,0002	**8,9570e-07**	**3,0804e-05**	0,0849	**1,364e-15**
Temperatura (T)	**1,9207e-11**	0,0337	0,0097	0,8537	**6,702e-13**
G×D	0,2221	0,0861	0,8716	0,6699	0,079525
G×T	0,0198	0,1427	0,9650	0,4513	0,003342
D×T	**5,9062e-07**	0,0120	**4,2747e-08**	0,9174	**4,769e-11**

Ryc. 6.11.A: punkty odpowiadające wysokiej temperaturze (czerwone i różowe) mają w ogólności wyższe wartości PC1 niż te odpowiadające optymalnej temperaturze (czarne i szare), jednocześnie dla pierwszego dnia (punkty czarne i czerwone) różnica jest znacząco większa niż dla dnia dziesiątego (punkty szare i różowe), co wskazuje na interakcję tych dwóch czynników. Punkty związane z pierwszym dniem mają na ogół wyższe wartości PC2 niż te związane z dniem dziesiątym.

Ryc. 6.11.B: w ogólności punkty związane z dniem pierwszym mają niższe wartości PC3 niż te związane z dniem dziesiątym. Można zauważyć również, że dla wysokiej temperatury różnica jest większa niż dla temperatury optymalnej.



Ryc. 6.11. Położenie prób w układzie współrzędnych głównych PC1, PC2 i PC3 otrzymanych na podstawie analizy PCoA dla podzbiorów skafoldów, dla *extracellular region*.

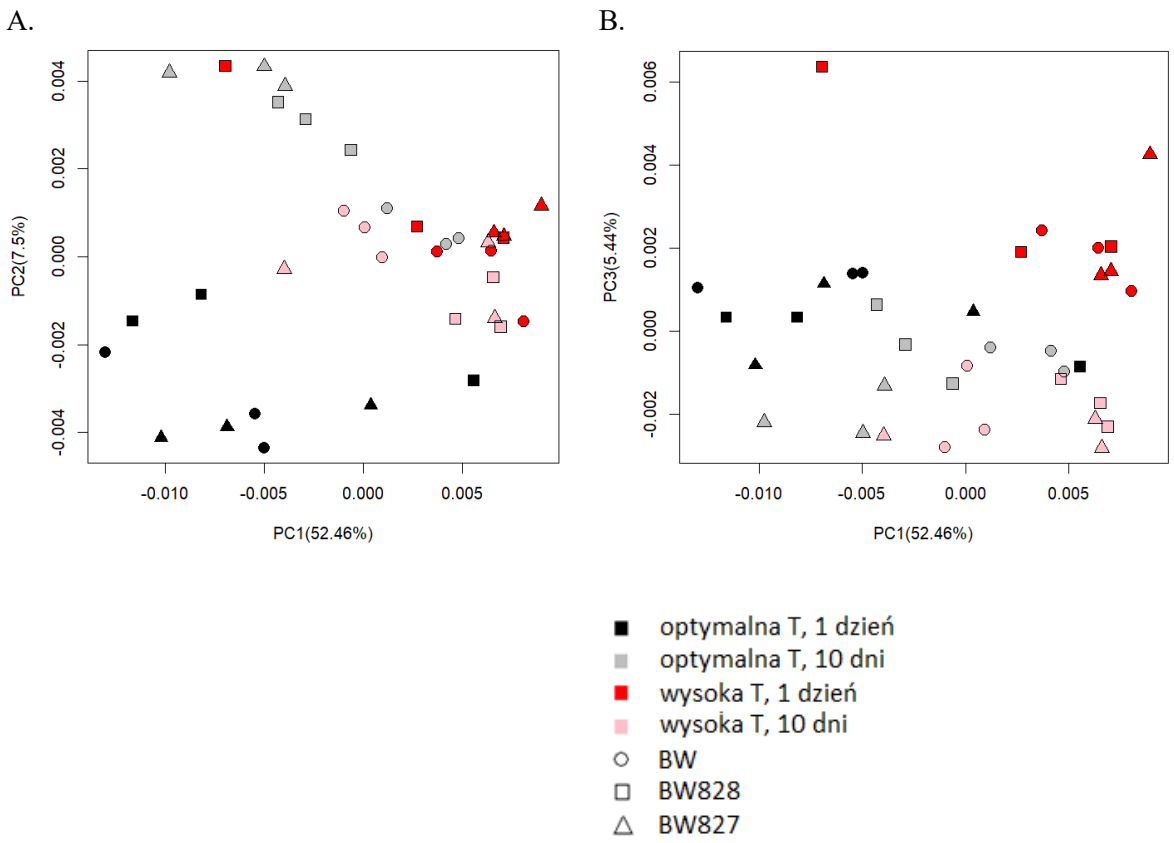
Przykład 3: *binding*

Tutaj również analiza została przeprowadzona jak w pierwszym przykładzie. Wykazany został związek położenia punktów względem osi PC1 z temperaturą, względem osi PC2 z dniem oraz interakcją czynników dzień i temperatura, natomiast względem osi PC3 tylko z dniem. Dla PC4 nie została wykazana istotność żadnego z czynników ani ich interakcji. Wyniki MANOVA były zgodne z wynikami ANOVA.

Tabela 6.14. Wyniki analizy wariancji dla podzbiorów *binding*.

Czynnik	Wartość P ANOVA				Wartość P MANOVA
	PC1 (52,46%)	PC2 (7,5%)	PC3 (5,44%)	PC4 (4,63%)	
Genotyp (G)	0,9332	0,0636	0,2114	0,7701	0,05776
Dzień (D)	0,4834	**4,6346e-06**	**1,1141e-08**	0,4567	**4,058e-12**
Temperatura (T)	**7,1856e-05**	0,3530	0,2171	0,0257	**2,644e-06**
G×D	0,2817	0,0449	0,6232	0,8762	0,11103
G×T	0,3527	0,1549	0,2355	0,4155	0,03706
D×T	0,0803	**5,8838e-09**	0,0002	0,1213	**4,788e-10**

Przygotowano wykresy ilustrujące położenie punktów odpowiadających zbiorom danych względem PC1, PC2 i PC3 (Ryc. 6.12).

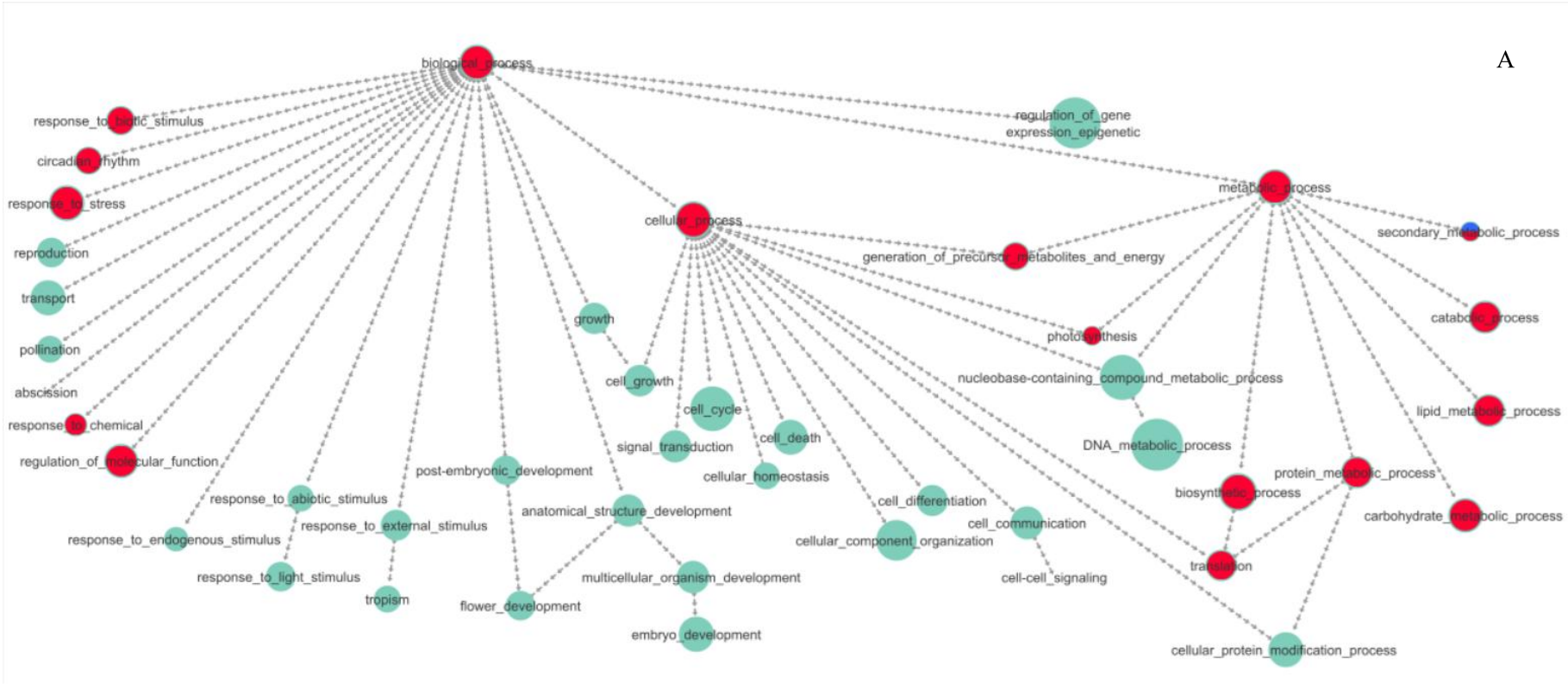


Ryc. 6.12. Położenie prób w układzie współrzędnych głównych PC1, PC2 i PC3 otrzymanych na podstawie analizy PCoA dla podzbiorów skafoldów, dla *binding*.

Ryc. 6.12.A: patrząc na oś PC1 można zauważyć, że punkty odpowiadające wysokiej temperaturze (czerwone i różowe) mają w ogólności wyższe wartości niż te odpowiadające optymalnej temperaturze (czarne i szare). Patrząc na PC2 możemy zaobserwować, że punkty odpowiadające pierwszemu dniowi (czerwone i czarne) mają w ogólności niższe wartości niż te odpowiadające dziesiątemu dniowi (różowe i szare), jednocześnie dla optymalnej temperatury to rozróżnienie jest wyraźniejsze niż dla wysokiej.

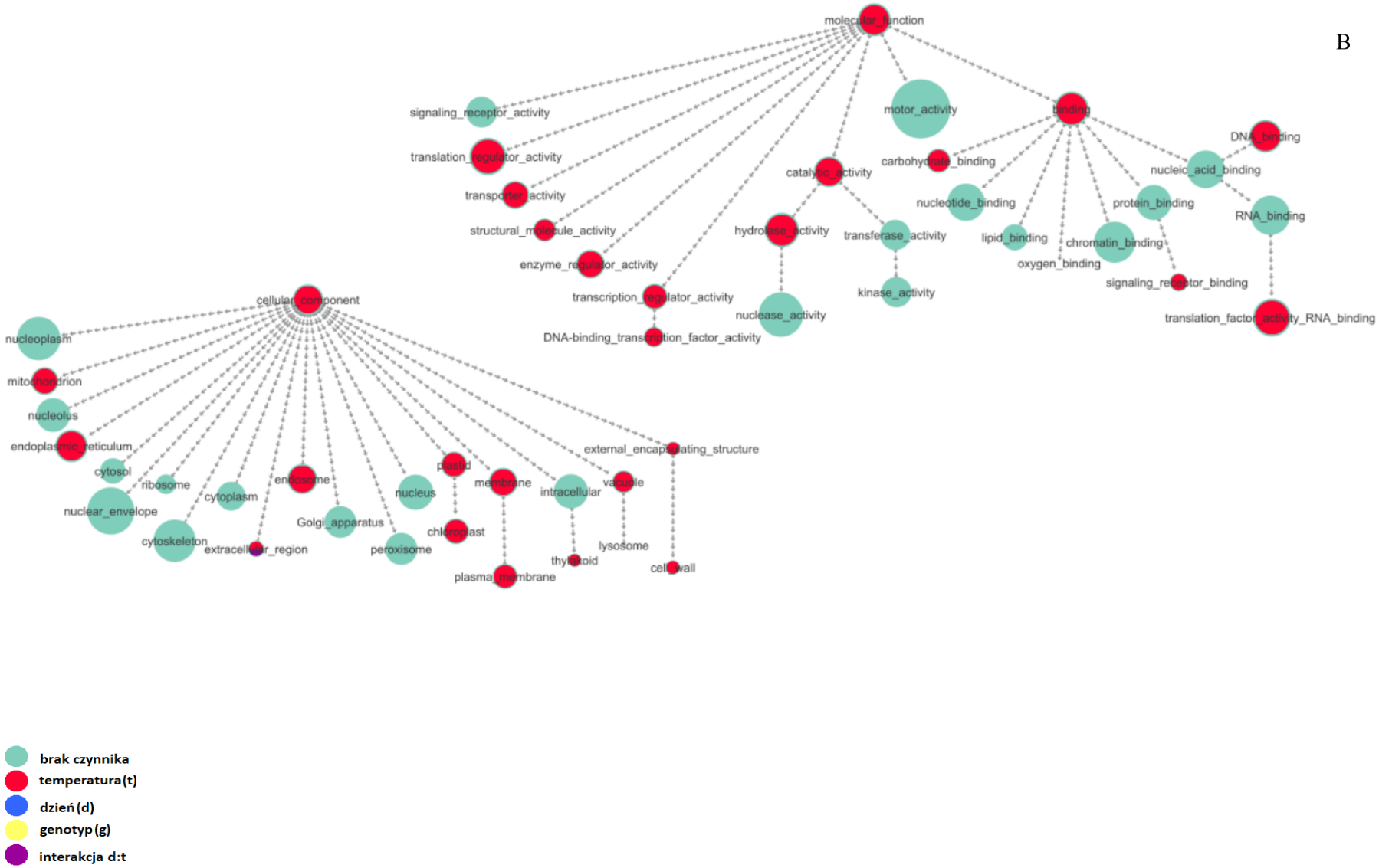
Ryc. 6.12.B: patrząc na PC3 widać, że punkty odpowiadające pierwszemu dniowi mają w ogólności wyższe wartości niż te odpowiadające dziesiątemu dniowi.

Zaproponujemy teraz jeszcze jeden sposób wizualnej prezentacji wyników analiz. Na rycinach (Ryc. 6.13.A-B, Ryc. 6.14.A-B, Ryc. 6.15.A-B i 6.16.A-B) pokazano grafy ontologii GO, zwizualizowane w programie Cytoscape (v. 3.8.1). Każdy graf obejmuje wyniki dla jednej z osi: PC1, PC2, PC3 i PC4, z podziałem na części ontologii opisujące: procesy biologiczne, funkcje molekularne i komponenty komórkowe. Wierzchołki grafu odpowiadają terminom z ontologii GOSlim, a krawędzie oznaczają relację "jest podklasą", co jest równoważne z zawieraniem się zbioru sekwencji dla podklasy w zbiorze dla klasy. Wielkość wierzchołków jest zależna od procentu zmienności objaśnianej przez daną współrzędną główną dla podzbiorów sekwencji wyznaczanych przez dany termin. Kolory wierzchołków zależą od czynników, które dla tych podzbiorów, w wyniku analizy wariancji, okazały się istotnie wpływać na zmienność sekwencji, czyli na położenie podzbiorów danych względem siebie. Znaczenie kolorów wierzchołków podają legendy rycin. Brak wierzchołka w postaci koła przy nazwie terminu GOSlim oznacza, że dla tego terminu nie znaleziono sekwencji w co najmniej jednym zbiorze danych, co skutkowało niekompletnością macierzy odległości Mash. Dotyczyło to czterech terminów: *abscission*, *cell-cell_signaling*, *lysosome* oraz *oxygen binding*.



brak czynnika
temperatura(t)
dzień(d)
genotyp(g)
interakcja d:t

B


molecular_function
signaling_receptor_activity
translation_regulator_activity
transporter_activity
structural_molecule_activity
enzyme_regulator_activity
transcription_regulator_activity
DNA-binding_transcription_factor_activity
motor_activity
catalytic_activity
hydrolase_activity
nuclease_activity
transferase_activity
kinase_activity
binding
carbohydrate_binding
nucleotide_binding
lipid_binding
oxygen_binding
chromatin_binding
protein_binding
signaling_receptor_binding
nucleic_acid_binding
DNA_binding
RNA_binding
translation_factor_activity_RNA_binding
cellular_component
nucleoplasm
mitochondrion
nucleolus
endoplasmic_reticulum
cytosol
ribosome
nuclear_envelope
cytoskeleton
cytoplasm
endosome
Golgi_apparatus
extracellular_region
peroxisome
nucleus
plastid
chloroplast
membrane
plasma_membrane
intracellular
thylakoid
vacuole
lysosome
external_encapsulating_structure
cell_wall
brak czynnika
temperatura(t)
dzień(d)
genotyp(g)
interakcja d:t

Ryc. 6.13. Graf ontologii GO dla PC1; A – procesy biologiczne; B – funkcje molekularne i komponenty komórkowe; kolor wierzchołka zależy od obecności istotnych efektów czynników lub ich interakcji.

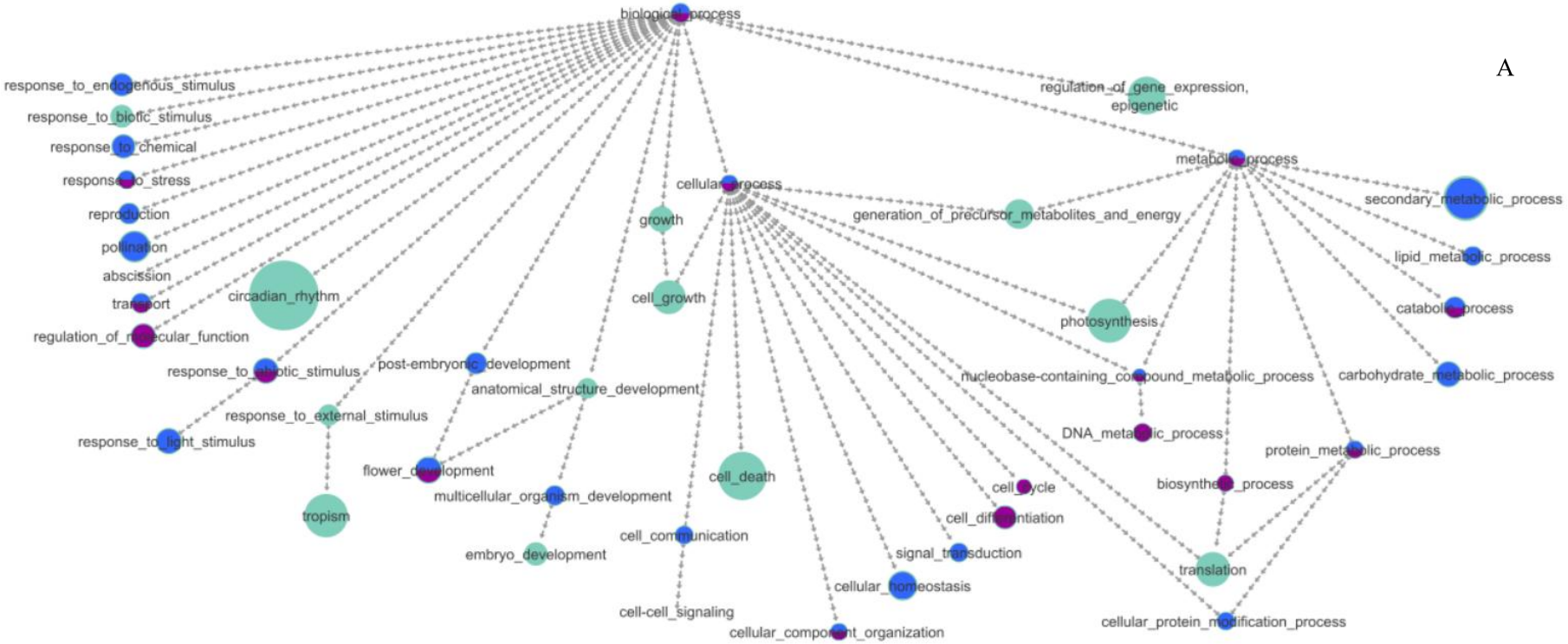
Ryciny 6.13.A i 6.13.B, dotyczące PC1, wskazują, że w 44 przypadkach istotny wpływ miał czynnik temperatura. Spośród tych przypadków: dla jednego *(secondary metabolic process*) istotny był również czynnik dzień, dla jednego (*extracellular region*) - interakcja dnia i temperatury. Dla 48 terminów analiza statystyczna nie wykazała istotności wpływu żadnego z czynników eksperymentalnych na zmienność danych sekwencyjnych. Analizując grafy możemy zaobserwować przypadki zanikania bądź pojawiania się istotnych czynników przy przejściu od klasy do podklasy.

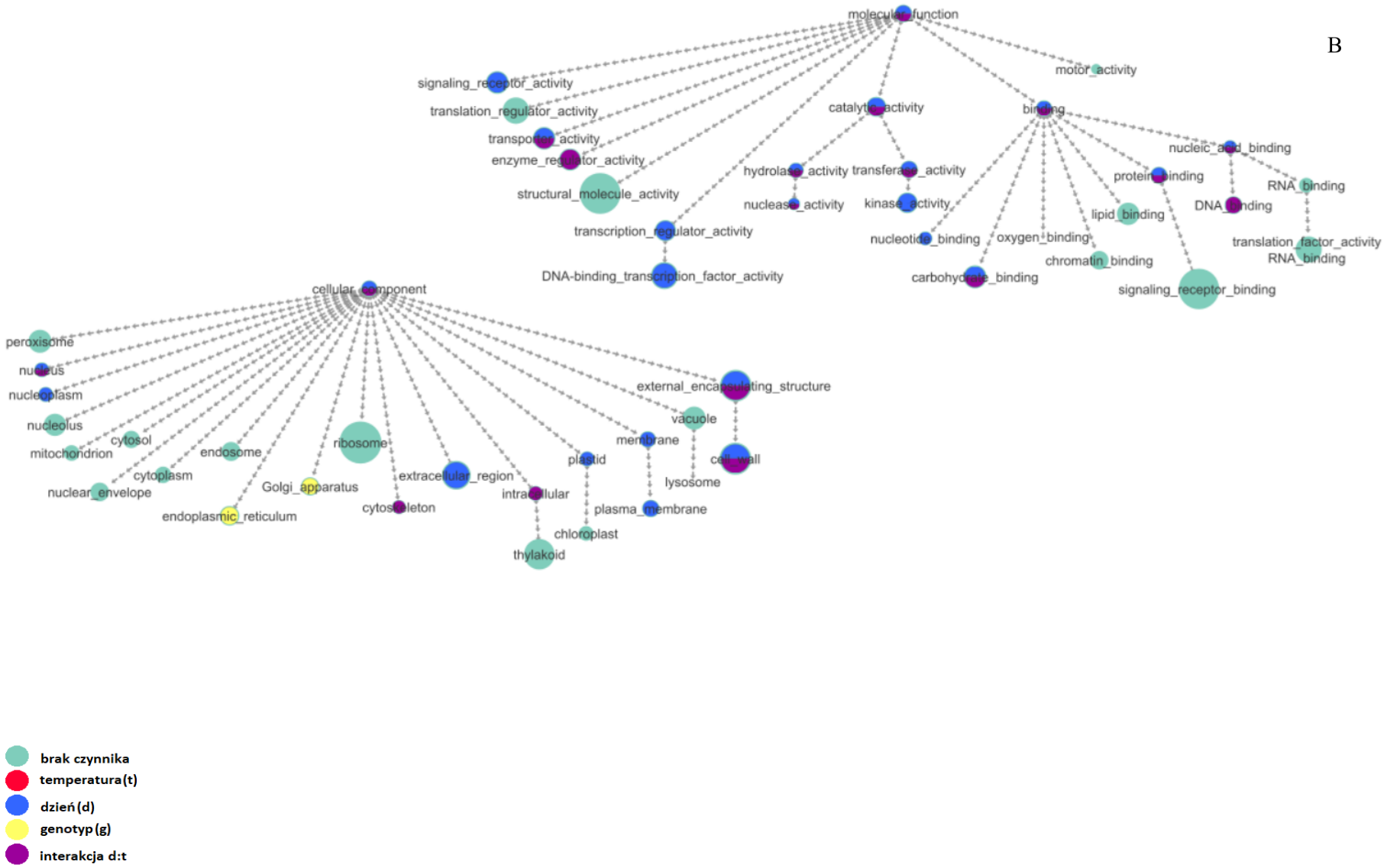
Dla terminu proces biologiczny oraz jego podklasy *metabolic process* istotny okazał się czynnik temperatura, natomiast dla terminu *secondary metabolic process* będącego podklasą *metabolic process* poza temperaturą pojawił się jeszcze jeden czynnik, mianowicie dzień. Oznacza to tyle, że jeśli z eksperymentalnych danych sekwencyjnych weźmiemy pod uwagę podzbiory przyporządkowane terminowi *biological process*, otrzymamy istotność czynnika temperatura i to samo stanie się przy rozważaniu zawartych w nich podzbiorów danych sekwencyjnych przyporządkowanych *metabolic process*, jednak gdy rozważamy jeszcze mniejszą podklasę – *secondary metabolic process* - i wyznaczane przez nią podzbiory danych sekwencyjnych, układ tych podzbiorów względem siebie różnicuje istotnie nie tylko temperatura, ale również dzień. Podobnie dla terminu *cellular component* istotnym czynnikiem była temperatura, natomiast dla podklasy *extracellular region* ważny był czynnik temperatura oraz interakcja temperatury z dniem.

W wielu przypadkach wystąpiło zanikanie istotności czynników eksperymentalnych – zaobserwowaliśmy to chociażby dla par klas: *cellular component* i *nucleoplasm*, *binding* i *lipid binding* czy też *cellular process* i *cell cycle*. W każdym z powyższych trzech przykładów dla większych podzbiorów, wyznaczanych przez klasy wymienione w pierwszej kolejności, analiza wariancji wykazała istotność czynnika temperatura, natomiast dla mniejszych podzbiorów, wyznaczanych przez podklasy tychże klas, istotność tego czynnika nie została wykazana.

Zaistniały również przypadki, gdy dla pewnej klasy dany czynnik był istotny, dla jej podklasy istotności nie wykazano, natomiast dla następnej podklasy zawartej w poprzednich dwóch dany czynnik znów był istotny. Dla PC1, w przypadku *molecular function*, znaleziono dwie takie trójki klas: *binding*, *nucleic acid binding* i *DNA binding* oraz *binding*, *protein binding* i *signaling receptor binding*.

A


biological_process
response_to_endogenous_stimulus
response_to_biotic_stimulus
response_to_chemical
response_to_stress
reproduction
pollination
abscission
transport
regulation_of_molecular_function
circadian_rhythm
response_to_abiotic_stimulus
response_to_external_stimulus
response_to_light_stimulus
tropism
post-embryonic_development
anatomical_structure_development
flower_development
multicellular_organism_development
embryo_development
cellular_process
growth
cell_growth
cell_death
cell_communication
cell-cell_signaling
cellular_component_organization
signal_transduction
cellular_homeostasis
cell_cycle
cell_differentiation
regulation_of_gene_expression,
epigenetic
metabolic_process
generation_of_precursor_metabolites_and_energy
photosynthesis
nucleobase-containing_compound_metabolic_process
DNA_metabolic_process
translation
cellular_protein_modification_process
biosynthetic_process
protein_metabolic_process
carbohydrate_metabolic_process
catabolic_process
lipid_metabolic_process
secondary_metabolic_process
brak czynnika
temperatura(t)
dzień(d)
genotyp(g)
interakcja d:t


B
molecular_function
signaling_receptor_activity
translation_regulator_activity
transporter_activity
enzyme_regulator_activity
structural_molecule_activity
transcription_regulator_activity
DNA-binding_transcription_factor_activity
motor_activity
catalytic_activity
hydrolase_activity
nuclease_activity
transferase_activity
kinase_activity
binding
nucleotide_binding
oxygen_binding
carbohydrate_binding
chromatin_binding
lipid_binding
protein_binding
signaling_receptor_binding
nucleic_acid_binding
DNA_binding
RNA_binding
translation_factor_activity
RNA_binding
cellular_component
peroxisome
nucleus
nucleoplasm
nucleolus
cytosol
mitochondrion
endosome
cytoplasm
nuclear_envelope
Golgi_apparatus
endoplasmic_reticulum
ribosome
cytoskeleton
extracellular_region
intracellular
thylakoid
plastid
chloroplast
membrane
plasma_membrane
external_encapsulating_structure
vacuole
cell_wall
lysosome
brak czynnika
temperatura(t)
dzień(d)
genotyp(g)
interakcja d:t

Ryc. 6.14. Graf ontologii GO dla PC2; A – procesy biologiczne; B – funkcje molekularne i komponenty komórkowe; kolor wierzchołka zależy od obecności istotnych efektów czynników lub ich interakcji.

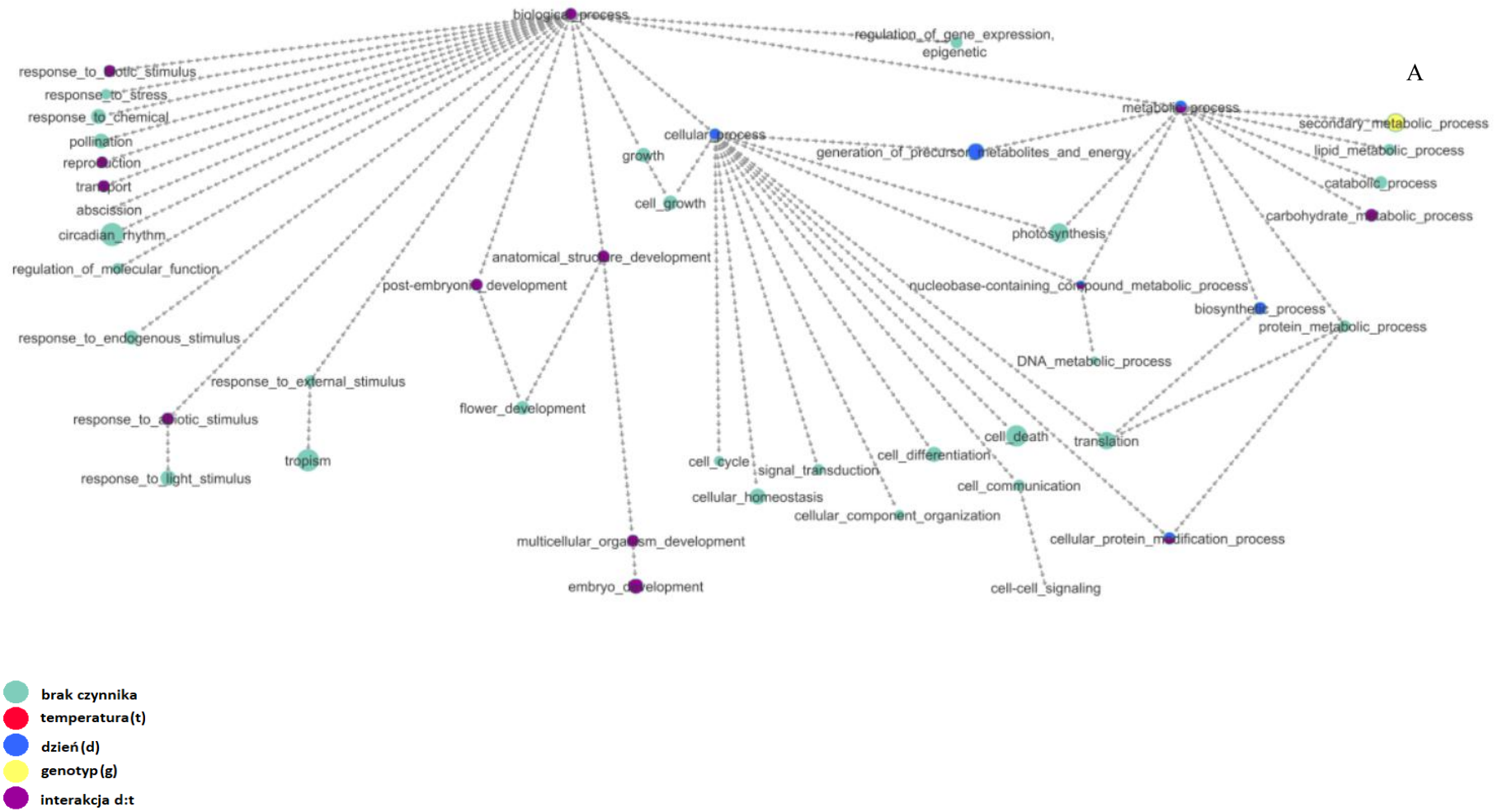
Ryciny 6.14.A i 6.14.B, dotyczące PC2, wskazują, że w 49 przypadkach istotny wpływ miał czynnik dzień. Spośród tych przypadków dla 25 terminów istotna była również interakcja dnia z temperaturą. Dla 9 terminów istotna była jedynie interakcja dnia z temperaturą. W dwóch przypadkach analiza wariancji wykazała istotność czynnika genotyp. Dla 32 terminów nie wykazano istotności żadnego z czynników eksperymentalnych. Podobnie jak dla PC1, na rycinach 6.13.A-B, dla PC2 można zaobserwować zanikanie, pojawianie się oraz zanikanie i pojawianie się istotnych czynników eksperymentalnych przy przejściu od klasy do podklasy.

Przykładem zanikania istotności czynnika przy przechodzeniu od klasy do podklasy była para terminów *nucleic acid binding* i *DNA binding* (Ryc. 6.14.B; funkcje molekularne). Dla *nucleic acid binding* istotnymi czynnikami eksperymentalnymi były dzień oraz interakcja dnia z temperaturą, natomiast dla *DNA binding* wykazana została tylko istotność czynnika dzień. Dla drugiej podklasy *nucleic acid binding*, mianowicie dla *RNA binding* nie wykazano istotności żadnego czynnika eksperymentalnego.

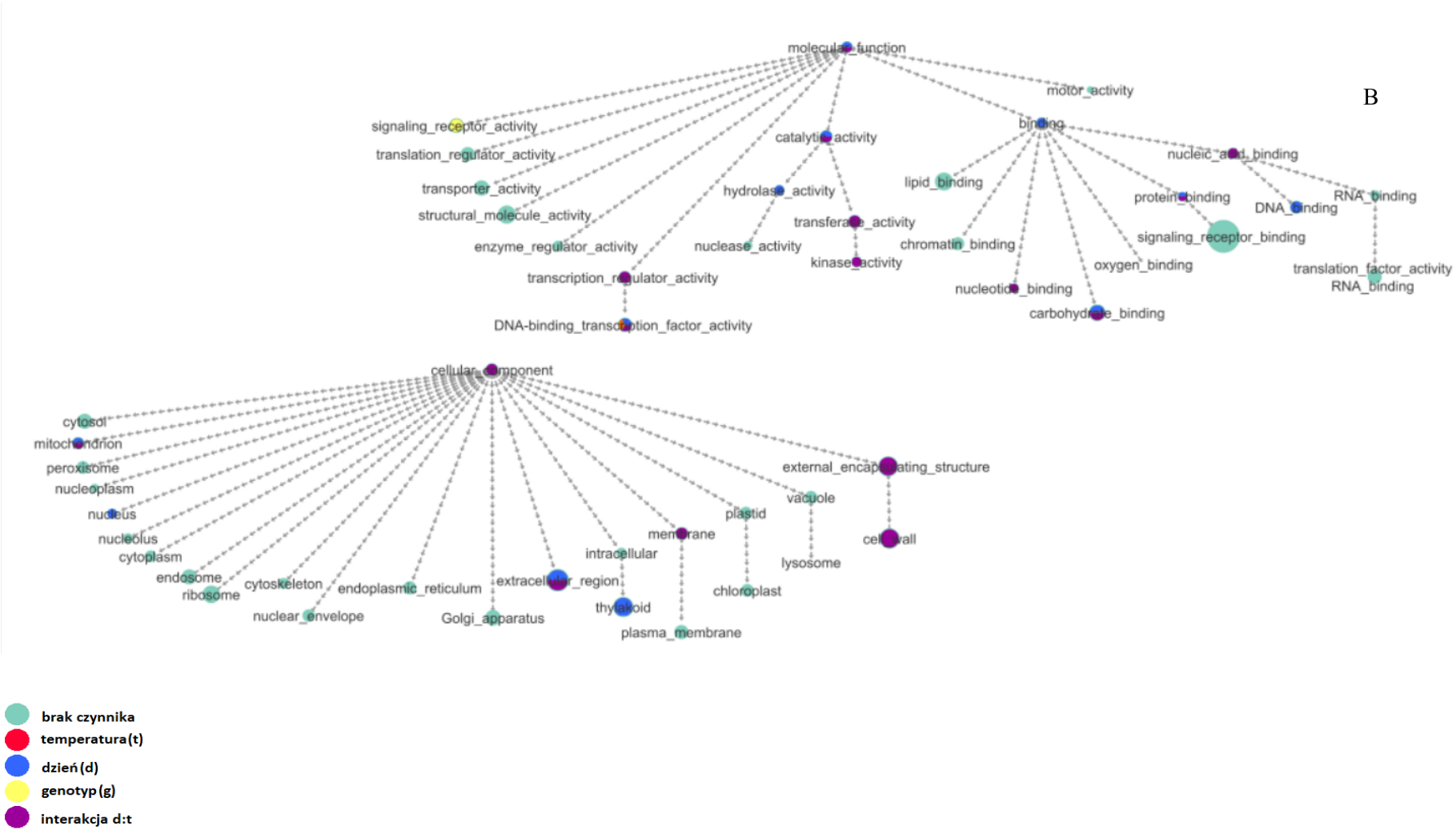
Przykładem zanikania i pojawiania się istotności w czasie przechodzenia z klasy do podklasy była trójka: *biological process*, *post-embryonic development* i *flower development*. Dla pierwszej klasy istotne były czynniki dzień i interakcja dnia z temperaturą, dla drugiej tylko czynnik dzień, a dla trzeciej znowu dzień oraz interakcja dnia z temperaturą.

Na rycinie 6.14.A dla PC2 można również zaobserwować układ, gdzie dla dwóch nadklas jednej klasy zbiory istotnych czynników są inne. Dzieje się tak na przykład dla klas *cellular process* i *biosynthetic process* oraz ich wspólnej podklasy *translation*: dla pierwszej klasy istotne były czynniki dzień oraz interakcja dnia z temperaturą, dla drugiej tylko interakcja dnia z temperaturą, natomiast dla ich podklasy żaden z czynników nie okazał się istotny. Innym przykładem są klasy *growth*, *cellular process* i ich podklasa *cell growth*, gdzie tylko dla *cellular process* wykazano istotność czynników dzień oraz interakcji dnia z temperaturą, natomiast dla dwóch pozostałych klas żaden czynnik nie okazał się istotny. Trzecim przykładem takiego układu jest trójka: *post-embryonic development*, *anatomical structure development* i *flower development* – tutaj z kolei dla podklasy wykazano więcej istotnych czynników niż dla jej dwóch nadklas, mianowicie dla *post-embryonic development* istotny okazał się czynnik dzień, dla *anatomical structure development* żaden z czynników nie był istotny, natomiast dla *flower development* istotny jest nie tylko czynnik dzień, ale i interakcja dnia z temperaturą.

Ciekawym przypadkiem są pary: *cellular component* i *Golgi apparatus* oraz *cellular component* i *endoplasmic reticulum*. Dla obu par dla nadklasy – *cellular component* – istotne były czynniki dzień oraz interakcja dnia z temperaturą, dla podklas żaden z tych czynników nie był istotny, natomiast istotność została wykazana dla czynnika genotyp.


A
biological_process
regulation_of_gene_expression,
epigenetic
response_to_biotic_stimulus
response_to_stress
response_to_chemical
pollination
reproduction
transport
abscission
circadian_rhythm
regulation_of_molecular_function
response_to_endogenous_stimulus
response_to_external_stimulus
response_to_abiotic_stimulus
response_to_light_stimulus
tropism
cellular_process
growth
cell_growth
anatomical_structure_development
post-embryonic_development
flower_development
multicellular_organism_development
embryo_development
cell_cycle
signal_transduction
cellular_homeostasis
cellular_component_organization
cell_differentiation
cell_death
cell_communication
cell-cell_signaling
metabolic_process
generation_of_precursor_metabolites_and_energy
photosynthesis
nucleobase-containing_compound_metabolic_process
DNA_metabolic_process
translation
cellular_protein_modification_process
biosynthetic_process
protein_metabolic_process
secondary_metabolic_process
lipid_metabolic_process
catabolic_process
carbohydrate_metabolic_process
brak czynnika
temperatura(t)
dzień(d)
genotyp(g)
interakcja d:t

B



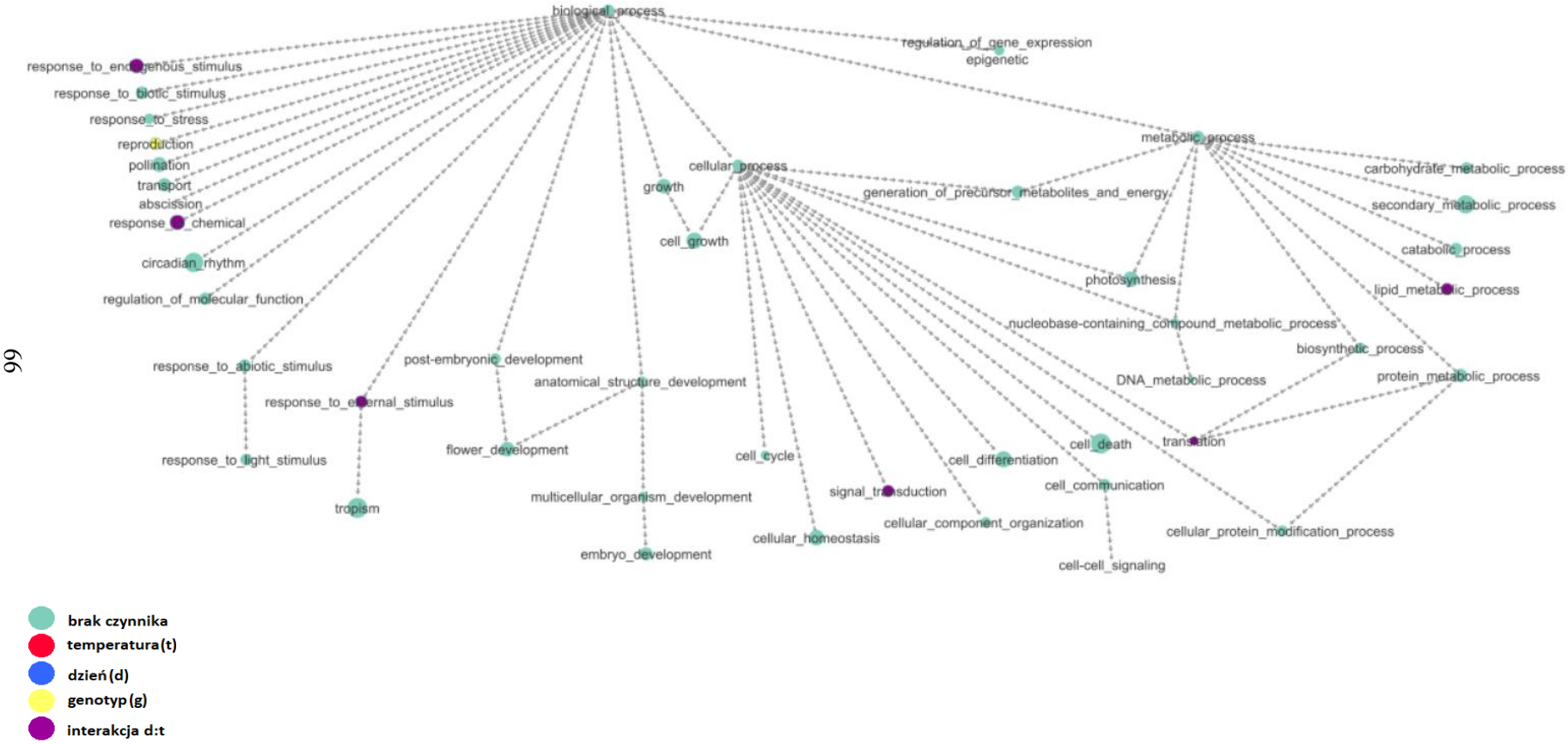
- brak czynnika
- temperatura(t)
- dzień(d)
- genotyp(g)
- interakcja d:t

Ryc. 6.15. Graf ontologii GO dla PC3; A – procesy biologiczne; B – funkcje molekularne i komponenty komórkowe; kolor wierzchołka zależy od obecności istotnych efektów czynników lub ich interakcji.

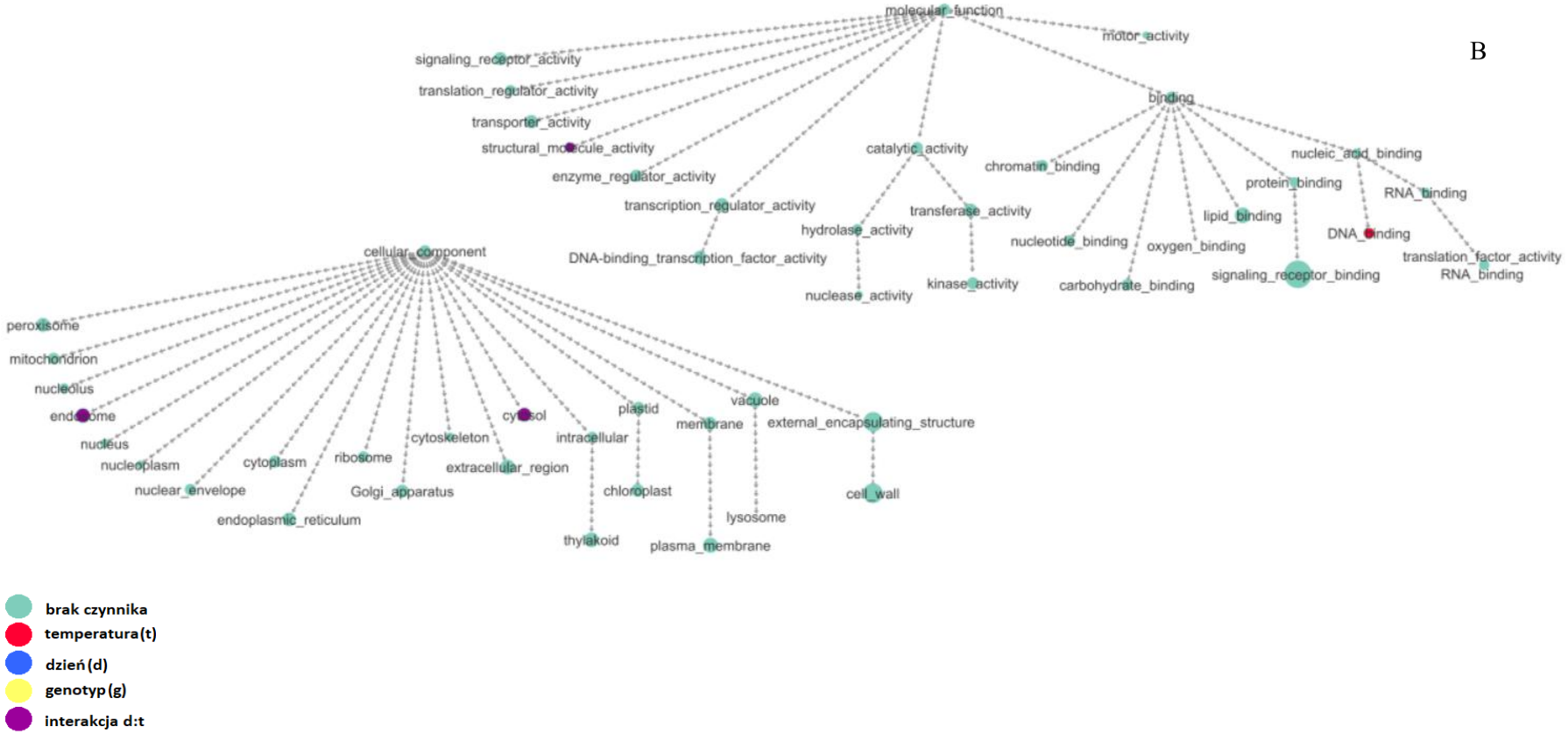
Ryciny 6.15.A-B, dotyczące PC3, wskazują, że dla 2 terminów istotny był genotyp. Dla 18 klas istotny był czynnik dzień – w tym dla 10 interakcja dnia z temperaturą i dla jednej dwie interakcje: dzień z temperaturą oraz genotyp z temperaturą. Dla 29 terminów istotna okazała się interakcja dnia z temperaturą. Dla 53 terminów żaden z czynników nie okazał się istotny. Podobnie jak dla poprzednich dwóch osi, również tu można zaobserwować pojawianie się istotnych czynników eksperymentalnych podczas przechodzenia z klas do podklas, na przykład dla *transcription regulator activity* istotnym czynnikiem była interakcja temperatury z dniem, natomiast dla *DNA-binding transcription factor activity* poza tą interakcją istotne były oba czynniki osobno. Zanikanie również występowało, na przykład dla *catalytic activity* istotnymi czynnikami były dzień oraz interakcja dnia z temperaturą, dla podklasy *transferase activity* tylko interakcja, a dla podklasy *hydrolase activity* – tylko czynnik dzień. Zanikanie i pojawianie się istotnych czynników jest kolejnym zjawiskiem pojawiającym się również dla przypadku PC3, czego przykładem jest trójka terminów: *biological process*, *cellular process* oraz *cellular protein modification process*, gdzie dla pierwszego i trzeciego terminu istotny był czynnik interakcji dnia z temperaturą (dla trzeciego również czynnik samego dnia), a dla drugiego tylko czynnik dnia. Co ciekawe, dla terminu *cellular protein modification process* inną nadklasą jest *protein metabolic process*, dla którego analiza nie wykazała istotności żadnego czynnika. Dla PC3 możemy również zaobserwować zanikanie istotności konkretnych czynników wraz z pojawieniem się zupełnie innego, czego przykładem jest para terminów: *metabolic process* (dzień i interakcja dnia z temperaturą) i *secondary metabolic process* (genotyp).

Ryciny 6.16.A-B, dotyczące PC4, wskazują na istotność genotypu dla jednego terminu, istotność temperatury dla jednego terminu oraz istotność interakcji dnia z temperaturą dla 9 terminów. Łącznie dla 81 terminów analiza wariancji nie wykazała istotności żadnego z czynników eksperymentalnych. Dla tej osi przykładem zanikania istotności czynników przy przechodzeniu z klasy do podklasy była para terminów *response to external stimulus* (interakcja czynnika dzień z czynnikiem temperatura) i *tropism* (brak czynników wpływających istotnie na odległości między zbiorami danych). Przykładem pojawiania się istotności była para terminów *biosynthetic process* (brak istotnych czynników) i *translation* (istotny czynnik interakcji dnia z temperaturą). Przykładem znikania i pojawiania się istotnych czynników była trójka terminów: *molecular function* (brak istotnych czynników), *catalytic activity* (istotny czynnik genotyp) i *transferase activity* (brak istotnych czynników).

A



brak czynnika
temperatura(t)
dzień(d)
genotyp(g)
interakcja d:t


B
molecular_function
motor_activity
signaling_receptor_activity
translation_regulator_activity
transporter_activity
structural_molecule_activity
enzyme_regulator_activity
transcription_regulator_activity
DNA-binding_transcription_factor_activity
catalytic_activity
hydrolase_activity
nuclease_activity
transferase_activity
kinase_activity
binding
chromatin_binding
nucleotide_binding
oxygen_binding
carbohydrate_binding
lipid_binding
protein_binding
signaling_receptor_binding
nucleic_acid_binding
DNA_binding
RNA_binding
translation_factor_activity
RNA_binding
cellular_component
peroxisome
mitochondrion
nucleolus
endosome
nucleus
nucleoplasm
nuclear_envelope
endoplasmic_reticulum
cytoplasm
ribosome
Golgi_apparatus
cytoskeleton
extracellular_region
cytosol
intracellular
thylakoid
plastid
chloroplast
membrane
plasma_membrane
vacuole
lysosome
external_encapsulating_structure
cell_wall
brak czynnika
temperatura(t)
dzień(d)
genotyp(g)
interakcja d:t

Ryc. 6.16. Graf ontologii GO dla PC4; A – procesy biologiczne; B – funkcje molekularne i komponenty komórkowe; kolor wierzchołka zależy od obecności istotnych efektów czynników lub ich interakcji.

6.3. Analiza różnicowa ekspresji genów uwzględniająca różnice transkryptomu

Ilościowa analiza ekspresji genów na podstawie danych RNA-seq polega najczęściej na badaniu istotności wybranych porównań pomiędzy wariantami doświadczalnymi uwzględnionymi w eksperymencie za pomocą metod bazujących na modelu analizy wariancji lub podobnych. W tej pracy podane są metody pozwalające na opisanie zmienności sekwencji transkryptów i porównanie wariantów pod względem samych sekwencji. Można oczekiwać, że szansa na ocenienie poziomu ekspresji w wybranych wariantach jako podobnego jest większa wtedy, gdy warianty te różnią się mało pod względem sekwencji, niż wtedy, gdy różnią się znacznie, gdyż w tym drugim przypadku na różnicę liczb zmapowanych odczytów wpływają także różnice w polimorfizmie transkryptów względem transkryptomu referencyjnego. Stąd powstaje pomysł, aby przeprowadzić ocenę istotności kontrastów w różnicowej analizie ekspresji genów z poprawką uwzględniającą podobieństwo sekwencji. Jeżeli w analizie różnicowej ekspresji stosuje się algorytm Deseq2, można to zrobić zastępując stosowany w nim zwykle model liniowy analizy wariancji modelem analizy kowariancji, dodając do modelu zawierającego czynniki doświadczalne zmienne objaśniające (towarzyszące) opisujące podobieństwo sekwencji. Za takie zmienne można przyjąć współrzędne główne wyliczone w analizie PCoA wykorzystującej macierz odległości Mash dla prób. Porównanie wyników uzyskanych w dwóch modelach powinno pozwolić na stwierdzenie, które geny wykazują zróżnicowany poziom ekspresji po uwzględnieniu poprawki na różnice sekwencji, inaczej mówiąc - po usunięciu wpływu podobieństwa sekwencji na wynik.

Zbiory danych NGS zostały zmapowane przy pomocy narzędzia TopHat (ver. 2.1.1.; Trapnell i in., 2009). Jako genom referencyjny posłużył genom *Hordeum vulgare* (wersja z dnia 07.04.2021). Następnie zastosowano algorytm Deseq2 do analizy kontrastów pomiędzy warunkami wysokiej temperatury a warunkami kontrolnymi, dla wszystkich genotypów obserwowanych w dniach 1 i 10.

W Tab. 6.15 przedstawione są wyniki analizy różnicowej ekspresji genów za pomocą algorytmu Deseq2 z wykorzystaniem dwóch modeli: modelu analizy wariancji (rozdz. 4.4, wzór 3) i modelu analizy kowariancji (wzór 4), przy czym w modelu analizy kowariancji za zmienne towarzyszące przyjęto wektory współrzędnych prób wyliczone w analizie PCoA dla danych Illumina.

Tabela 6.15. Liczby genów o ekspresji zróżnicowanej pomiędzy warunkami wysokiej temperatury i warunkami kontrolnymi według algorytmu Deseq2 z modelem analizy wariancji i z modelem analizy kowariancji.

Genotyp	Kierunek zmiany ekspresji genów	Analiza wariancji		Analiza kowariancji	
		Dzień obserwacji		Dzień obserwacji	
		1	10	1	10
Bowman	W dół	1111	28	955	25
	W górę	252	267	203	238
	Razem	1363	295	1158	263
BW827	W dół	889	338	832	289
	W górę	439	228	385	199
	Razem	1328	566	1217	488
BW828	W dół	510	351	469	312
	W górę	216	210	176	192
	Razem	726	561	645	504

Liczby genów o zróżnicowanej ekspresji (DEG, ang. diiferentially expressed genes) były na ogół mniejsze w dniu 10 niż 1 zarówno w przypadku analizy wariancji, jak i dla analizy kowariancji – za wyjątkiem genotypu Bowman, liczby DEG o kierunku „w górę" oraz BW828, dla analizy kowariancji, dla zmiany ekspresji „w górę". Dla wszystkich wariantów doświadczalnych, dla których oceniano kontrasty, liczby DEG uzyskane w wyniku analizy kowariancji były mniejsze niż uzyskane w wyniku analizy wariancji – zachodziło to dla obu kierunków zmiany ekspresji. Tak więc dodanie zmiennych opisujących podobieństwo sekwencji transkryptów do modelu analizy kontrastów zmniejszyło liczbę znalezionych DEG. Inaczej - można powiedzieć, że nie jest wykluczone, iż pewna część DEG znalezionych w analizie wariancji została uznana za geny o zróżnicowanej ekspresji dlatego, że sekwencje ich transkryptów różniły się pomiędzy warunkami OT i HT.

7. Dyskusja

W tej pracy sformułowany został problem „jakościowej analizy porównawczej transkryptomów roślin w doświadczeniach wieloczynnikowych” i zaproponowano odpowiednie dla niego metody analizy danych (odległość Mash, PCoA, ANOVA). W tej sekcji zajmiemy się zagadnieniami związanymi z metodami i otrzymanymi wynikami, które nie zostały jeszcze poruszone, oraz rozwiniemy te, które do tej pory zostały jedynie zasygnalizowane.

Pierwszą rzeczą, nad którą warto się głębiej zastanowić, jest odległość Mash jako podstawa całej analizy. Stosowana odległość nie jest jedyną możliwą opcją - wybór powinien być dokonany przy uwzględnieniu zapotrzebowania, specyfiki badanej zmienności oraz dostępnych zasobów ludzkich, czasowych i obliczeniowych. Jakie są zatem argumenty przemawiające za wyborem odległośći Mash w przypadku opisanym w tej pracy?

Odległość Mash pozwala znaleźć różnice między transkryptami, o których była mowa w sekcji 2.2 ("Natura i badanie zmian sekwencji nukleotydowych"). Z definicji, odległość Mash wykrywa SNP, insercje, delecje, rearanżacje itp. Takie różnice wykrywa też odległość Hamminga, jednakże ta druga może być stosowana tylko dla sekwencji równych długości. Ponadto, wpływ insercji i delecji na odległość Hamminga jest znacznie większy niż wpływ SNP, jak pokazano na przykładzie w sekcji 2.3.2, za wyjątkiem sekwencji powtarzalnych. Dalej, odległość Mash jest dobrze zdefiniowana jako odległość zbiorów sekwencji, co udowodniono w sekcji 2.3.3, podczas gdy dla odległości Hamminga należałoby podać dodatkową definicję umożliwiającą porównywanie zbiorów.

Wybór odległości Mash jest również uzasadniony zamiarem badania różnic jakościowych, a nie ilościowych, pomiędzy transkryptomami. W obliczeniach k-mery są wykorzystywane tylko jeden raz, co czyni wynik niezależnym od powtórzeń sekwencji (k-merów). Ta własność związana jest z początkowym zastosowaniem odległości Mash do badania różnic pomiędzy genomami. W takim przypadku może być ona interpretowana w terminach liczby mutacji przy założeniu modelu Poissona. W przypadku analizy transkryptomów, taka interpretacja nie wydaje się właściwa – dlatego proponujemy nieco inne spojrzenie: dzięki modelowi Poissona jest to funkcja określona na zbiorze transkryptomów, spełniająca warunki na odległość i pozwalająca wykorzystać indeks Jaccarda do określania położenia zbiorów danych względem siebie, co otwiera możliwość porównywania zbiorów w kontekście jakościowym.

Wybór odległości Mash jest też korzystny ze względu na koszt czasowy obliczeń. Dla odległości Damareau-Levenshteina (Zhao i Sahni, 2020) pomiędzy dwiema sekwencjami a i b złożoność obliczeniowa algorytmu liczącego jej wartość wynosi O(|a|*|b|) (gdzie |a| i |b| to odpowiednio długości sekwencji a i b, a symbol O oznacza oszacowanie górne), co znaczy, że dla dwóch genomów długości n będzie to $O(|n|^2)$. Oczekiwany czas pracy programu wyznaczającego dwa skecze potrzebne do obliczenia odległości Mash pomiędzy dwoma genomami o długości n wyraża się wzorem O(2n + 2 s log s log n)], gdzie *s* oznacza rozmiar skeczu. Wyznaczenie skeczu jest najbardziej czasochłonnym procesem, a estymacja indeksu Jaccarda to koszt rzędu O(*s*), stąd też czas obliczeń dla algorytmu Mash - określony przez funkcję niemal liniową - jest mniejszy niż czas potrzebny do wyliczenia odległości Damareau-Levenshteina, który określony jest przez funkcję kwadratową.

Algorytm Mash wybiera losowo s *k*-merów z każdego ze zbiorów sekwencji. Dla rzędu *k* istnieje 4^k różnych *k*-merów, czyli na przykład istnieją 1024 różne 5-mery, a różnych 10-merów jest już ponad

milion. Zasadne jest więc pytanie, jaki rozmiar *k*-merów oraz zbiorów *k*-merów reprezentujących zbiory danych pozwala na dobre przybliżenie ich rzeczywistego indeksu Jaccarda bez ponoszenia dużych kosztów obliczeń. Wybór tych parametrów został szczegółowo opisany w artykule Ondova i in. (2016). Mniejsze rozmiary *k*-merów pozwalają na oszczędność pamięci, jednak są mniej czułe na różnice między zbiorami sekwencji – nie tylko im bliższe filogenetycznie są porównywane organizmy, tym większa szansa, że wiele *k*-merów będą miały wspólnych, ale również dla dwóch losowych genomów rośnie szansa znalezienia dużego zbioru wspólnych *k*-merów. Z kolei większe rozmiary *k*-merów pozwalają na lepsze rozróżnienie bardzo podobnych zbiorów sekwencji, jednak wymagają więcej pamięci obliczeniowej oraz są mniej odporne na błędy występujące w zbiorach sekwencji. Wybór rozmiaru skeczu *s* warunkuje zbudowanie dobrej reprezentacji zbiorów danych ale też czas wykonywania obliczeń. Granica błędu przybliżenia indeksu Jaccarda wyraża się wzorem $\varepsilon = O(\frac{1}{\sqrt{s}})$, zatem zależy tylko od parametru *s*. Parametry *k*=21 i *s*=1000 zostały sprawdzone w wyżej wymienionej pracy pod kątem dokładności estymacji i działały dobrze dla wybranych przypadków genomów (w tym metagenomów). W niniejszej pracy odległości między zbiorami danych zostały policzone kilkukrotnie, a wynikowe macierze odległości nie różniły się pomiędzy sobą.

Warto jeszcze przypomnieć o kwestii reprezentacji w zbiorach *k*-merów tylko leksykograficznie mniejszych (występujących wcześniej w porządku alfabetycznym) z par komplementarnie odwrotnych *k*-merów, w przypadku zaistnienia tychże. Takie postępowanie pozwala uniknąć dwukrotnej reprezentacji jednej grupy *k*-merów, kiedy inne mogły nie wystąpić w parze w zbiorze danych uzyskanym w wyniku sekwencjonowania i przetwarzania danych i wskutek tego otrzymać tylko jednokrotną reprezentację.

Obliczenia prowadzone były przy użyciu komputerów z Poznańskiego Centrum Superkomputerowo-Sieciowego. Przy takich możliwościach sprzętowych obliczenia trwały stosunkowo krótko, a ponadto wykorzystywano możliwość zrównoleglenia dużej liczby zadań, np. przy liczeniu macierzy odległości. Przykładowo, dla zbiorów surowych danych BW828OT1dR3 i BW828HT1dR1 czas wykonywania komendy mash, czyli odległości pomiędzy parą zbiorów danych, wynosił: całkowity czas pomiędzy wywołaniem i zakończeniem działania programu (real time) – 8 minut 20,360 sekund; czas, w którym procesor wykonywał pracę (user time) – 7 minut 5,981 sekund; czas który kod programu spędził w jądrze systemu operacyjnego (sys time) – 0 minut 4,622 sekund. Dla skafoldów czas wynosił odpowiednio: real time – 0 minut 4,675 sekund, user time – 0 minut 3,899 sekund, sys time – 0 minut 0,027 sekund. Występowały zatem znaczne różnice w czasie wykonywania obliczeń dla danych surowych i złożonych, na rzecz skafoldów.

Innym zagadnieniem związanym z różnicami wielkości danych jest interpretacja odległości. W artykule Ondova i in. (2016) wspomniane jest, że należy być ostrożnym przy porównywaniu genomów o znacznych różnicach w długościach. W naszym przypadku transkryptomy będące podstawą zbiorów danych pochodzą z organizmów bardzo blisko spokrewnionych. Można przyjąć, że różnice wielkości zbiorów danych wynikające z różnej liczby odczytów dla różnych prób, które w praktyce są akceptowane w analizie danych, nie wpływają na interpretację wyników. Różnica między wielkością surowych danych a tymi po procesie składania (patrz sekcja 6.1.2) też nie powinna mieć znaczenia, ponieważ liczone przez odległość Mash zbiory *k*-merów będą różne tylko w tych miejscach, w których surowe dane zawierają sekwencje, które przy składaniu zostały odrzucone jako błędne lub nigdzie nie pasujące. W przypadku, gdyby porównaniu podlegały próby pochodzące z różnych gatunków, np. roślinnych, efekt wielkości zbiorów danych może wystąpić, jakkolwiek nie jest prawdopodobnie duży ze względu na ograniczone wahania liczb genów u różnych gatunków.

Zastanówmy się nad przypadkiem, gdy porównywane zbiory danych są parami rozłączne w tym sensie, że nie mają wspólnych k-merów. W zastosowaniach rozważanych w tej pracy taki przypadek jest mało prawdopodobny, gdyż mówimy na ogół o doświadczeniach, w których badane są blisko spokrewnione organizmy, a więc zbiór wspólnych k-merów nie jest pusty. W kontekście roślin, nawet w przypadku odległych gatunków, wspólne k-mery wystąpią najprawdopodobniej dla genów ortologicznych. Zauważmy również, że przy pustym zbiorze wspólnych k-merów, czyli wartości współczynnika Jaccarda równej 0, wartość odległości Mash jest nieokreślona. Natomiast algorytm zastosowany w programie liczącym odległość Mash nadaje w tej sytuacji odległości wartość 1. Zapewnia to spójność otrzymywanych wyników.

Program liczący odległość Mash pomija k-mery zawierające nieodczytane nukleotydy, to znaczy: jeśli w przesuwanym oknie czytającym kolejne k-mery pojawi się taki zawierający obserwację „N”, to nie zostaje mu przypisany hasz i nie trafia w tej postaci do szkicu. Stąd zbiór k-merów reprezentujących porównywane zbiory sekwencji zawiera tylko w pełni odczytane fragmenty. Gdyby tak nie było, algorytm musiałby przypisywać wartości "N" jakąś informację aby porównać zawierający ją k-mer z *k*-merami innego zbioru. W związku z tym musiałaby następować jedna z dwóch sytuacji: *k*-mer otrzymałby oddzielny hasz i byłby odczytywany jako element wspólny dwóch zbiorów tylko w przypadku wystąpienia braku odczytu dokładnie w tym samym miejscu w tym drugim zbiorze (co zakłada, że spodziewamy się braku odczytu w tym miejscu również dla drugiego zbioru danych), albo musiałby otrzymać oznaczenie korespondujące z haszami dla k-merów o identycznych odczytach w innych miejscach niż „N” oraz dowolnym odczycie w miejscu „N”, co prowadziłoby do utożsamiania ze sobą tychże k-merów. Prowadziłoby to do sytuacji, w której odległość Mash zawierałaby informację o jakości sekwencjonowania kosztem informacji o różnicach między sekwencjami.

W pracy przedstawiono wyniki obliczeń dla trzech zestawów zbiorów danych, pochodzących z różnych etapów przetwarzania: danych NGS, kontigów i skafoldów. Pokazano bardzo duże podobieństwo wyników obliczeń na zbiorach kontigów i zbiorach skafoldów i do interpretacji wybrano tylko drugi z tych wariantów. Wyniki otrzymane na podstawie danych NGS i na podstawie skafoldów różniły się między sobą, co wymaga dyskusji.

Odległości liczone na podstawie danych NGS różniły się w sensie "położenia" ich rozkładu znacznie od odległości liczonych na skafoldach - były większe. Wynika to z tego, że w NGS jest większa różnorodność k-merów niż w skafoldach, w których pominięte są k-mery nie pasujące do innych w sensie grafu De Bruijna. To powoduje, że dla danych NGS iloraz liczby wspólnych k-merów do liczby wszystkich k-merów jest mniejszy. K-mery powstające w skafoldach podczas składania sekwencji prawdopodobnie nie wpływają na wynik porównania przy założeniu, że sekwencjonowanie było przeprowadzone z dostateczną głębokością.

W sekcji 6.1 zauważono, że wyniki analizy wariancji wykonanej na danych surowych i skafoldach były podobne. Biorąc pod uwagę czas potrzebny do wykonania obliczeń można by sądzić, że bardziej korzystne jest wykonanie analizy na skafoldach. Jednak ten wniosek nie bierze pod uwagę kosztu wykonania tychże skafoldów. Także pomijając aspekt czasu obliczeń, wydaje się, że korzystniejsza jest analiza skafoldów, które prawdopodobnie stanowią lepszą reprezentację transkryptomu w sensie jakościowym niż dane surowe. W sytuacji, gdy nie jest możliwe złożenie danych surowych w skafoldy, wyniki dla danych surowych mogą stanowić przybliżenie wyników, które uzyskano by dla skafoldów.

Interpretacja istotności efektów dla poszczególnych współrzędnych głównych w analizie wariancji również wymaga omówienia. Analiza współrzędnych głównych miała na celu jak najlepiej przybliżyć rozmieszczenie względem siebie danych wielowymiarowych w przestrzeni o niższym wymiarze.

Skupiliśmy się na czterech pierwszych współrzędnych. Dla przedstawionych wykresów Sheparda, ilustrujących zgodność między odległościami rzeczywistymi, a tymi uzyskanymi dla pierwszych sześciu wymiarów w analizie współrzędnych głównych, dodawanie kolejnych wymiarów po czwartym nie poprawiało w sposób znaczący wyników regresji liniowej; dla surowych danych wynik regresji liniowej wskazywał na znacznie słabszy liniowy związek między odległościami Mash a odległościami przedstawionymi przez PCoA niż wynik regresji liniowej dla skafoldów, dla których dawała zadowalające wyniki. Suma procentu zmienności uchwyconego przez cztery kolejne wymiary dla danych surowych była równa 28,94% (z piątym wymiarem: 32,84%), a dla skafoldów była równa 71,51% (z piątym wymiarem: 75,61%). Można uznać, że pierwsze cztery wymiary z analizy współrzędnych głównych dobrze przybliżają odległości Mash między zbiorami danych dla skafoldów. Jak wytłumaczyć różnicę pomiędzy stopniem wyjaśniania zmienności transkryptomów i dopasowaniem regresji liniowej odległości pomiędzy analizą prowadzoną na danych NGS i na skafoldach? Wydaje się, że głównie tym, iż dane NGS zawierają więcej szumu i błędów sekwencjonowania niż skafoldy, ponadto skafoldy są skonstruowane z k-merów metodą grafów de Bruijna. Z tego wynika większa wymiarowość oryginalnej przestrzeni, w której znajdują się zbiory danych NGS, choć dokładnego wymiaru tej przestrzeni, przy obecnej wiedzy, nie możemy podać. Natomiast wiadomo, że wymiarowość przestrzeni transformowanej za pomocą analizy współrzędnych głównych ma duży wpływ na stopień wyjaśniania zmienności przez pierwsze wyznaczone współrzędne.

Mimo różnic w stopniu objaśniania zmienności i w wynikach regresji dla wykresów Sheparda, wyniki dalszych analiz dla surowych danych i skafoldów były niemal identyczne. Podobieństwo to wynika z faktu, że analiza PCoA ma za zadanie raczej uchwycić zmienność między danymi niż faktyczne relacje odległości. Miejscem, gdzie dwie analizy się różniły, był wynik analizy wariancji dotyczący istotności czynnika genotyp i interakcji genotyp × temperatura. Dla tych przypadków istotność w analizie danych NGS miała miejsce dla trzeciej lub czwartej współrzędnej, które wyjaśniały małą część zmienności.

ANOVA umożliwia sprawdzenie, czy dane czynniki lub ich interakcje wpływają istotnie na zadaną zmienną, natomiast MANOVA pozwala to sprawdzić dla kilku zmiennych naraz. W tym przypadku zmiennymi były współrzędne punktów odzwierciedlające odległości zbiorów danych względem siebie. Stosując ANOVA sprawdzono istotność czynników względem kolejnych wymiarów, natomiast używając MANOVA sprawdzono to samo względem położenia punktów w przestrzeni czterowymiarowej. W przypadkach, gdy MANOVA potwierdzała istotność efektów, których nie wskazała ANOVA, można uznać, że dopiero analiza w czterowymiarowej przestrzeni pozwoliła wychwycić różnice, które nie były widoczne dla poszczególnych wymiarów i stwierdzić istotność zadanego czynnika lub interakcji czynników. W przeciwnym przypadku, to znaczy gdy ANOVA potwierdza istotność czynników niepotwierdzonych przez MANOVA, przyczyn można doszukiwać się w tym, że dla pojedynczych wymiarów różnice mogły wydawać się większe niż były dla czterech wymiarów. Ogólnie, wyniki MANOVA były zgodne z analizą poszczególnych współrzędnych. Pokazały one, że głównymi źródłami zmienności są dzień i temperatura oraz interakcja tych dwóch czynników. Wykonanie obu analiz, jedno- i wielowymiarowej, pozwoliło, z jednej strony, na potwierdzenie wyników, z drugiej strony, na uzyskanie szerszego obrazu zmienności zbiorów danych.

Żeby uzyskać bardziej szczegółowe wyniki dotyczące wpływu czynników eksperymentalnych na rozmieszczenie zbiorów sekwencji względem siebie, rozbiliśmy je na podzbiory odpowiadające terminom GOSlim i przeprowadziliśmy dla każdego z nich analogiczną analizę jak dla całych zbiorów danych. Jest to punkt, w którym znaleźliśmy się najbliżej interpretacji biologicznej wyników – w innych miejscach, mimo wykazania pewnych zależności między odległościami a czynnikami eksperymentalnymi, nie zostaje poczynione żadne głębsze wnioskowanie poza prostymi

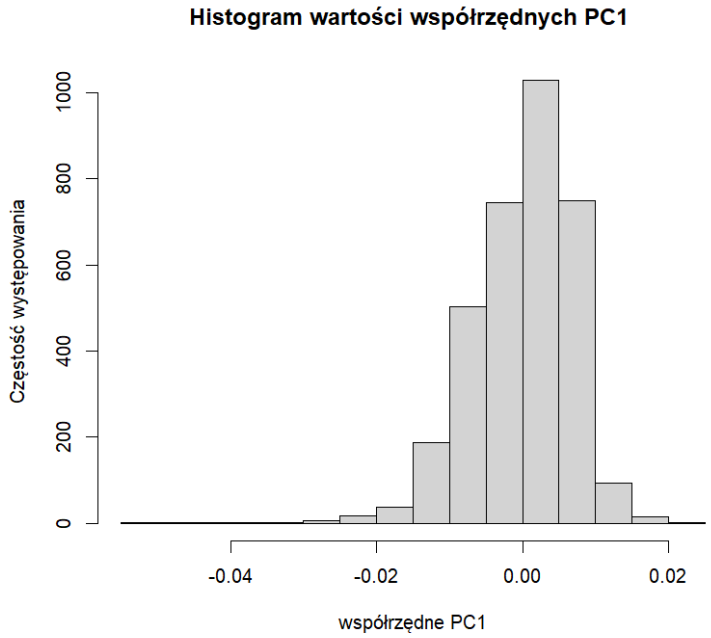
stwierdzeniami, które wynikają bezpośrednio z obliczeń. Natomiast ontologia GOSlim pozwoliła nam połączyć wyniki analiz z sensem biologicznym. Wybór tej ontologii został dokonany głównie ze względu na stosunkowo niewielki rozmiar i – co za tym idzie – prostsze i szybsze obliczenia. Oczywiście, jak już zostało wspomniane wcześniej (rozdział 4.6), istnieje wiele ontologii i w każdym przypadku prowadzenia analizy z użyciem tego typu informacji warto dobrać taką, która najlepiej odpowiada zarówno tematowi badań, jak i poziomowi szczegółowości, jaki zamierza się osiągnąć. Warto w tym miejscu nadmienić, że wiarygodność informacji zawartych w wykorzystywanym referencyjnym zbiorze transkryptów zadnotowanych względem ontologii również może być do pewnego stopnia kontrolowana poprzez dobór tylko tych adnotacji, które znalazły się tam na podstawie doniesień z zaufanych źródeł (np. potwierdzone eksperymentalnie, wyniki zamieszczone w opublikowanym artykule). W przypadku rozważanego tutaj zbioru danych, analiza wykorzystująca anotację GOSlim pozwoliła m.in. na stwierdzenie, iż efekt interakcji dnia i temperatury, nie uznany za istotny w przypadku ogólnej analizy wariancji dla pierwszej współrzędnej głównej, okazał się jednak istotny dla terminu GOSlim *extracellular region*, co pokazuje praktyczną przydatność proponowanej analizy.

Przy podziale zbiorów danych na podzbiory wyznaczone przez kategorie GOSlim, dla pewnych kategorii nie we wszystkich zbiorach danych znalezione zostały sekwencje powiązane z nimi (za pomocą blast). Przyczyną tego jest brak w zbiorze danych sekwencji dostatecznie podobnych do transkryptów referencyjnych związanych z tymi kategoriami. To z kolei może wynikać z niskiej lub zerowej ekspresji pewnych genów lub z pojawienia się transkryptów (izoform) nieuwzględnionych w danych referencyjnych bądź też z błędów sekwencjonowania. Wyjście z tej sytuacji poprzez np. zmianę czułości w mapowaniu programem blast nie jest zalecanym rozwiązaniem. Z definicji, odległość Mash zbioru pustego od dowolnego zbioru sekwencji jest nieokreślona. Arbitralne ustalenie odległości między pustymi zbiorami danych a resztą zbiorów mogłoby przekłamać wyniki. Sposobem zastosowanym w tej pracy było wykonanie pełnych analiz tylko dla niepustych podzbiorów i załączenie informacji, które zbiory danych zostały pominięte z tytułu braku danych. W związku z powyższym, przy liczeniu macierzy odległości pominięte zostały wiersze i kolumny związane z pustymi podzbiorami, a PCoA i ANOVA zostały wykonane na niekompletnym zbiorze prób. Takie rozwiązanie ma sens przy braku niewielkiej ilości danych. Należy dodać, że być może istniałby sposób na uzupełnienie brakujących danych w macierzy odległości Mash za pomocą metod stosowanych w statystyce. W takim przypadku, należałoby stosować rozwiązania dedykowane dla sytuacji występowania braków nielosowych, zależnych od poziomu ekspresji danego genu lub od dokładności sekwencjonowania (Carpenter i Smuk, 2021).

W pracy opisane również zostało wykorzystanie wyników z analizy PCoA jako zmiennych objaśniających do analizy różnicowej poziomu ekspresji genów. W wyniku dodania w modelu informacji z analizy współrzędnych głównych we wszystkich kontrastach zmniejszyła się liczba genów uznanych za ulegające istotnej zmianie ekspresji. Oznacza to, że odległości te objaśniają do pewnego stopnia różnice w ekspresji genów. Zatem, dodanie informacji o odległościach może okazać się pomocne, jeśli chcemy ocenić różnice w ekspresji genów z poprawką na podobieństwa transkryptomu. Badaniem, które nie zostało przeprowadzone w ramach tej rozprawy, ale również mogłoby przynieść ciekawe wyniki, jest przyjrzenie się odległościom między sekwencjami odpowiadającymi tym genom, które utraciły status zróżnicowanej ekspresji (to znaczy tym, których zmienność ekspresji jest objaśniana przez odległości). Można by wtedy sprawdzić, czy bliskie/dalekie odległości miały bezpośredni wpływ na analizę różnicową ekspresji genów, czy też nie jest to tak bezpośrednie przełożenie na wyniki.

Jest to postępowanie w pewnym sensie podobne do stosowanego w analizie asocjacyjnej (GWAS). W tym wypadku, analizuje się cechę ilościową obserwowaną w populacji akcesji w modelu liniowym zawierającym polimorfizm wybranego SNP oraz, jako dodatkowe zmienne objaśniające, współrzędne akcesji w przestrzeni PCoA wyznaczone na podstawie dużej liczby polimorfizmów, co pozwala na poprawienie (adjustment) efektu analizowanego SNP ze względu na podobieństwo akcesji ocenione za pomocą dużej liczby pozycji w genomie (podobieństwo sekwencji genomowych). Określa się to mianem „poprawienia ze względu na strukturę populacji".

Następną kwestią, którą należy poruszyć w dyskusji, jest normalność danych wykorzystywanych w analizie wariancji. Na Ryc.7.1 przedstawiono histogram wartości współrzędnych PC1 zebranych z analiz PCoA dla wszystkich terminów GOSlim. Taki wektor wartości ma rozmiar odpowiedni do sprawdzania normalności rozkładu. Test Shapiro-Wilka odrzucił hipotezę o normalności rozkładu na poziomie $P < 0{,}001$. Jak widać, odrzucenie hipotezy wynika głównie ze skośności rozkładu – występuje tutaj lewoskośność, przy współczynniku skośności -0.8735. Literatura wskazuje, że odchylenia od normalności danych związane z ich skośnością są przy takim współczynniku skośności stosunkowo dobrze tolerowane przez ANOVA (Blanca i in., 2017). W przypadku danych o rozkładzie odbiegającym od normalnego, jednym ze sposobów stosowanych w analizie wariancji jest obliczanie poziomów istotności poprzez permutacje danych. Takie obliczenia wykonano dla pierwszych trzech współrzędnych głównych w analizie skafoldów. Otrzymane wyniki nie różniły się istotnie od wyników uzyskanych poprzez zwykłą analizę wariancji, co dodatkowo wskazuje na to, że w przypadku tych danych skośność nie zaburza prowadzonych analiz w sposób nakazujący odrzucić tę metodę na rzecz innych narzędzi statystycznych.



Ryc.7.1. Histogram wartości współrzędnych PC1 zebranych z analiz PCoA dla wszystkich terminów GOSlim.

Wyniki analiz wykazały istotność wpływu czynników temperatura i dzień na odległości między transkryptomami, a czynnik genotyp na ogół nie był istotny. Znaczy to, że metoda nie wykazała

istotnych statystycznie różnic w kontekście jakościowym, co może wskazywać, że te rośliny, reprezentujące różne genotypy, mają bardzo podobne lub nawet takie same mechanizmy obronne, a różnią się dopiero ilościowo w ekspresji.

Warto w dyskusji jeszcze odnieść się do wyników analiz uzyskanych na tych samych danych w pracy (Mikołajczak i in., 2022). Ponieważ przeprowadzone zostały one pod kątem konkretnego pytania badawczego, dobrane metody różnią się znacząco od tej przedstawionej w niniejszej pracy, co jednak nie wyklucza przeprowadzenia pewnych porównań.

Podobieństwo badanych linii do genotypu Bowman ocenione w publikacji na podstawie SNPów trzech typów (frakcja markerów typu RNA-seq, 50k Chip, GBS) wyniosło: BW827 – 70,12%, BW828 – 10,44%, co wskazuje na większe podobieństwo BW827 do BW niż BW828. Ta analiza bazowała na próbach uzyskanych w warunkach kontrolnych w 1d (trzy połączone powtórzenia). Wynik ten został potwierdzony odległością Mash. Średnie odległości pomiędzy trzema powtórzeniami dla OT, d1 wynosiły: BW-BW827: 0,0094, BW-BW828: 0,0112.

Wspólnym punktem analiz zawartych w artykule Mikołajczaka i in. (2022) i w niniejszej pracy wydaje się być analiza różnicowa ekspresji genów. Tutaj jednak również konieczne jest wskazanie miejsc, gdzie metody w tych pracach się różniły oraz komentarz, jakie ma to konsekwencje w kontekście interpretacji i porównywania wyników. Przypomnijmy, że w tej rozprawie analiza DEG została przeprowadzona na dwa sposoby – z poprawką na odległość sekwencji i bez poprawki. W artykule nie wzięto pod uwagę DEG polimorficznych (poza tymi związanymi z sdw1 – znajdującymi się w prawdopodobnym regionie delecji tego genu). Porównując wyniki, najwięcej DEG znaleźliśmy w analizie bez poprawek, co jest oczywiste w przypadku artykułu, a w przypadku przedstawionym w tej pracy zostało już omówione. Zarówno przy zastosowaniu analizy kowariancji jak i odrzuceniu polimorfizmów, DEG było mniej wszędzie za wyjątkiem jednego przypadku: dla odmiany Bowman w 10 dniu wynik był taki sam. Wniosek: stosując różne metody na usunięcie wpływu w różnicy sekwencji na analizę różnicową ekspresji genów uzyskano podobne wyniki, tzn. redukcję liczby DEG podobnego rzędu.

W artykule szukano także DEG, dla których zmiana poziomu ekspresji w reakcji na temperaturę była różna w różnych momentach czasowych - nazwano je DRG (differentially reacting genes). Zauważmy, że w niniejszej pracy wykazano, że interakcja czynników czas i temperatura była istotna statystycznie w analizie jakościowej transkryptów. Można postawić pytanie, czy analizę DRG, która bazuje na analizie kontrastów podobnie jak analiza DEG, można "poprawić" podobnie jak powyżej analizę DEG stosując analizie kowariancji w Deseq2. Pozostaje to do wyjaśnienia, podobnie jak to, czy przedstawione tutaj wyjaśnienie zmienności transkryptomów za pomocą adnotacji funkcjonalnej GOSlim ma odniesienie do konkretnych polimorfizmów znalezionych w analizie przedstawionej przez Mikołajczaka i in. (2022).

Wreszcie, ogólny wniosek z artykułu jest taki, że istniało więcej DEG w porównaniach pomiędzy HT i OT niż w porównaniach pomiędzy genotypami - biorąc pod uwagę poziom ekspresji genów. W niniejszej pracy również czynnik temperatura okazał się ważniejszy niż czynnik genotyp - biorąc pod uwagę różnice w sekwencjach transkryptów.

Wnioski

1. Opracowana metoda pozwala na analizę sprawdzającą powiązanie między różnicami jakościowymi transkryptomów (RNA) uzyskanych w wyniku doświadczeń wieloczynnikowych, a uwzględnionymi czynnikami eksperymentalnymi; jest to możliwe dzięki wykorzystaniu odległości Mash.
2. Przedstawiony dowód tego, że odległość Mash spełnia warunki bycia odległością w sensie matematycznym, przemawia na korzyść stosowania tej funcji w analizie danych doświadczalnych.
3. Zaprezentowane własności odległości Mash, Hamminga oraz Damareau-Levenshteina wskazują, że wybór odległości Mash do analizy jakościowych różnic pomiędzy transkryptomami był właściwy.
4. Opracowana metoda analizy danych transkryptomicznych jest stosunkowo mało wymagająca pod względem czasu i zasobów potrzebnych do obliczeń.
5. Zastosowanie odległości Mash oraz ontologii GOSlim pozwoliło na wzbogacenie analizy danych transkryptomicznych o biologiczny aspekt funkcjonalny.
6. Pokazano, że opracowana metoda pozwala na zmodyfikowanie różnicowej analizy ilościowej ekspresji genów w ten sposób, że brane są pod uwagę podobieństwa sekwencji transkryptomów w poszczególnych zbiorach danych, co może być właściwe w przypadku niektórych zastosowań.
7. Metoda została sprawdzona na rzeczywistym zbiorze danych transkryptomicznych, pozyskanych w wyniku eksperymentu wieloczynnikowego.
8. Porównując wyniki analiz różnicowych ekspresji genów i analizy odległości Mash bazującej na różnicach sekwencji transkryptów zauważono podobieństwa i różnice zastosowanych podejść i metod obliczeniowych.

(nr)	Próba	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	BWOT1dR1	0	0,0107	0,0063	0,0094	0,011	0,0107	0,0138	0,012	0,0104	0,0095	0,01	0,0109	0,0087	0,0081	0,0074	0,0105
2	BWOT1dR2	0,0411	0	0,0105	0,0158	0,0184	0,0181	0,0217	0,0195	0,018	0,0147	0,0163	0,0156	0,0107	0,0143	0,0114	0,0135
3	BWOT1dR3	0,0399	0,0419	0	0,0099	0,012	0,0116	0,0146	0,0129	0,0112	0,0104	0,0107	0,0119	0,0081	0,0086	0,0071	0,0113
4	BWOT10dR1	0,0419	0,0429	0,0402	0	0,0056	0,0058	0,0095	0,0079	0,0073	0,0073	0,0074	0,009	0,0129	0,008	0,0105	0,0098
5	BWOT10dR2	0,0455	0,048	0,0436	0,0403	0	0,0043	0,0068	0,0062	0,0059	0,0072	0,0083	0,0092	0,0153	0,0078	0,0124	0,0114
6	BWOT10dR3	0,0436	0,0452	0,0411	0,0398	0,0388	0	0,0067	0,0065	0,0059	0,0071	0,0084	0,0087	0,0153	0,0076	0,012	0,0114
7	BWHT1dR1	0,0438	0,045	0,0426	0,04	0,0404	0,0381	0	0,0057	0,0058	0,0096	0,0104	0,0116	0,0191	0,0098	0,0155	0,0156
8	BWHT1dR2	0,0413	0,0433	0,0421	0,0391	0,0373	0,0394	0,035	0	0,0052	0,008	0,0088	0,0099	0,0169	0,009	0,0135	0,0136
9	BWHT1dR3	0,0426	0,0439	0,0403	0,0411	0,039	0,0382	0,036	0,0354	0	0,0079	0,0083	0,0096	0,0147	0,0076	0,0114	0,0124
10	BWHT10dR1	0,0435	0,0404	0,0411	0,04	0,0441	0,0435	0,0406	0,0399	0,0416	0	0,0071	0,0071	0,0129	0,0082	0,0111	0,0101
11	BWHT10dR2	0,0441	0,0457	0,0413	0,0398	0,0439	0,0427	0,0409	0,0394	0,04	0,0435	0	0,0087	0,0134	0,0092	0,0115	0,0114
12	BWHT10dR3	0,041	0,0411	0,0406	0,0396	0,0406	0,0417	0,0387	0,0371	0,0394	0,0353	0,0417	0	0,0139	0,0095	0,0118	0,011
13	BW827OT1dR1	0,0395	0,0394	0,0387	0,0414	0,0441	0,0419	0,0406	0,0413	0,0411	0,0411	0,0436	0,0386	0	0,0114	0,0073	0,012
14	BW827OT1dR2	0,0402	0,0403	0,0388	0,0406	0,0436	0,0427	0,0413	0,0404	0,0406	0,0394	0,0426	0,0369	0,0376	0	0,0088	0,011
15	BW827OT1dR3	0,0414	0,0413	0,0391	0,0407	0,0435	0,0417	0,042	0,0399	0,0416	0,0414	0,0438	0,0395	0,039	0,0381	0	0,0119
16	BW827OT10dR1	0,0421	0,0402	0,0406	0,0404	0,046	0,0438	0,0435	0,0442	0,0447	0,0381	0,0433	0,0391	0,0377	0,039	0,0414	0
17	BW827OT10dR2	0,0409	0,041	0,0392	0,0388	0,0452	0,0447	0,0449	0,0442	0,0452	0,0398	0,0447	0,0395	0,039	0,0382	0,0414	0,0368
18	BW827OT10dR3	0,0398	0,0406	0,0386	0,0372	0,042	0,041	0,0413	0,0402	0,0427	0,0383	0,0416	0,0363	0,0396	0,0372	0,0396	0,0366
19	BW827HT1dR1	0,055	0,0544	0,0533	0,0498	0,0477	0,051	0,0473	0,0482	0,0472	0,0527	0,0433	0,0498	0,0523	0,0515	0,051	0,0519
20	BW827HT1dR2	0,045	0,0446	0,0439	0,041	0,0404	0,0429	0,0377	0,0376	0,0379	0,0421	0,0417	0,0402	0,041	0,0417	0,0424	0,0452
21	BW827HT1dR3	0,0457	0,0463	0,0427	0,0419	0,0419	0,0421	0,0387	0,0392	0,039	0,0392	0,0436	0,0394	0,0438	0,0398	0,0444	0,0419
22	BW827HT10dR1	0,0424	0,045	0,043	0,0402	0,0416	0,0414	0,0382	0,0394	0,0402	0,0373	0,0419	0,0355	0,0419	0,0399	0,0423	0,0427
23	BW827HT10dR2	0,0493	0,0467	0,0457	0,0447	0,0444	0,0457	0,0455	0,0438	0,0452	0,0444	0,036	0,0419	0,047	0,0439	0,0457	0,045
24	BW827HT10dR3	0,047	0,0457	0,0444	0,0435	0,0417	0,0411	0,0399	0,0395	0,0409	0,0416	0,0376	0,0381	0,0441	0,043	0,0438	0,045
25	BW828OT1dR1	0,0403	0,0395	0,0407	0,0432	0,0452	0,0426	0,0438	0,0432	0,0433	0,039	0,0452	0,039	0,039	0,0378	0,0409	0,039
26	BW828OT1dR2	0,0407	0,0416	0,0414	0,0402	0,0463	0,0432	0,047	0,0438	0,0447	0,0416	0,0436	0,0407	0,0403	0,0407	0,0404	0,041
27	BW828OT1dR3	0,0432	0,0482	0,0439	0,0417	0,0421	0,04	0,0387	0,0398	0,0386	0,0444	0,0423	0,0403	0,0433	0,0404	0,0423	0,0454
28	BW828OT10dR1	0,043	0,043	0,041	0,0382	0,0411	0,0421	0,0417	0,0413	0,0424	0,0395	0,0435	0,0384	0,0407	0,0386	0,0421	0,0391
29	BW828OT10dR2	0,0413	0,041	0,0414	0,0406	0,0411	0,0403	0,0407	0,0392	0,041	0,0388	0,0426	0,0353	0,0381	0,0391	0,0414	0,0386
30	BW828OT10dR3	0,0452	0,0429	0,0442	0,0438	0,0441	0,046	0,0455	0,0433	0,0455	0,0406	0,0427	0,0392	0,0432	0,0419	0,0433	0,0414
31	BW828HT1dR1	0,0472	0,0449	0,0439	0,0409	0,0416	0,0409	0,0372	0,039	0,0382	0,0402	0,0426	0,0384	0,0433	0,0391	0,0435	0,0421
32	BW828HT1dR2	0,0407	0,0402	0,0417	0,0388	0,0407	0,0407	0,0366	0,0383	0,0367	0,0388	0,0417	0,0373	0,0395	0,0392	0,0399	0,0374
33	BW828HT1dR3	0,0416	0,0416	0,0417	0,0414	0,0426	0,0426	0,0411	0,0403	0,041	0,0387	0,0436	0,0399	0,0409	0,0419	0,0399	0,0398
34	BW828HT10dR1	0,0417	0,0409	0,0399	0,0395	0,0406	0,0426	0,0387	0,039	0,039	0,0368	0,0413	0,0371	0,0384	0,0388	0,0417	0,0376
35	BW828HT10dR2	0,0429	0,0463	0,0406	0,0423	0,04	0,0392	0,0369	0,0363	0,0372	0,0403	0,0394	0,0373	0,042	0,0402	0,0423	0,0447
36	BW828HT10dR3	0,0496	0,0473	0,047	0,0454	0,0435	0,0446	0,0424	0,0433	0,0452	0,043	0,0394	0,0413	0,0465	0,045	0,047	0,0477

Tabela S1: Macierze odległości Mash dla danych NGS i skafoldów; szary – dane NGS, biały – skafoldy; wartości zaokrąglone do czwartego miejsca po przecinku;

(nr)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	0,0114	0,0094	0,021	0,0117	0,0122	0,0127	0,0151	0,0134	0,0088	0,0103	0,0108	0,0101	0,0101	0,0095	0,0129	0,0111	0,0111	0,0112	0,0134	0,016
2	0,0125	0,0134	0,0286	0,0188	0,0191	0,0192	0,0198	0,0199	0,011	0,0114	0,0184	0,0154	0,0138	0,0132	0,0198	0,0164	0,0133	0,018	0,021	0,0224
3	0,0119	0,0099	0,0216	0,0121	0,013	0,0132	0,0152	0,0138	0,0088	0,0099	0,0116	0,0113	0,0109	0,0104	0,0135	0,011	0,011	0,012	0,0134	0,0161
4	0,0127	0,0085	0,0167	0,0079	0,0081	0,0081	0,015	0,0086	0,0114	0,0136	0,0084	0,0071	0,0077	0,0081	0,009	0,0082	0,0111	0,0079	0,0087	0,0116
5	0,0151	0,0098	0,0152	0,006	0,0066	0,0067	0,0161	0,0072	0,0125	0,0152	0,0072	0,0077	0,0092	0,0099	0,0073	0,0077	0,0127	0,0071	0,0071	0,0104
6	0,0158	0,0095	0,0149	0,0059	0,0064	0,0064	0,016	0,0067	0,013	0,0157	0,0068	0,0074	0,0092	0,0101	0,0066	0,0067	0,0122	0,0066	0,0068	0,0094
7	0,0198	0,0131	0,014	0,0055	0,0059	0,0065	0,0193	0,0082	0,0154	0,0182	0,0071	0,0111	0,0125	0,0138	0,0051	0,0082	0,0145	0,0075	0,0072	0,0105
8	0,0166	0,0114	0,0138	0,0053	0,0058	0,0068	0,0173	0,0075	0,0136	0,0161	0,0068	0,0096	0,0112	0,0121	0,006	0,0076	0,0122	0,0073	0,0069	0,0101
9	0,016	0,0103	0,0139	0,0052	0,006	0,0068	0,0161	0,0079	0,0121	0,0144	0,0063	0,009	0,0099	0,0106	0,0056	0,0065	0,0116	0,0071	0,0072	0,0102
10	0,0136	0,0088	0,0172	0,0081	0,0088	0,008	0,0136	0,0082	0,0106	0,0133	0,009	0,0076	0,0086	0,0086	0,0091	0,0083	0,0113	0,0078	0,0082	0,011
11	0,0139	0,0098	0,0162	0,0091	0,0094	0,009	0,0128	0,0086	0,011	0,0134	0,0094	0,0084	0,0092	0,0097	0,0098	0,0096	0,0118	0,0082	0,0087	0,0114
12	0,0128	0,0098	0,0184	0,0106	0,0109	0,0094	0,0134	0,0091	0,011	0,0135	0,0107	0,0096	0,0099	0,0101	0,0114	0,01	0,0118	0,0091	0,0103	0,0128
13	0,0115	0,0118	0,0266	0,0162	0,017	0,0162	0,0177	0,0174	0,0095	0,0085	0,0154	0,0129	0,0114	0,0114	0,0175	0,0145	0,0128	0,0158	0,0168	0,0202
14	0,0145	0,0101	0,0176	0,0085	0,0089	0,0085	0,0163	0,0096	0,0103	0,0126	0,0074	0,0086	0,0096	0,0101	0,0085	0,0081	0,0123	0,0085	0,0088	0,0117
15	0,0114	0,0099	0,0228	0,0129	0,0137	0,0135	0,0155	0,0142	0,0083	0,008	0,0121	0,0112	0,0105	0,0103	0,014	0,0124	0,0114	0,0125	0,0136	0,0169
16	0,0095	0,0083	0,0221	0,0127	0,013	0,0128	0,0161	0,0128	0,011	0,013	0,0133	0,0086	0,0089	0,0086	0,0144	0,0111	0,0114	0,0128	0,0145	0,0165
17	0	0,0099	0,026	0,0171	0,018	0,0177	0,0177	0,0173	0,011	0,0105	0,0174	0,0116	0,0114	0,0108	0,0185	0,0158	0,0118	0,0164	0,0184	0,0206
18	0,0354	0	0,0198	0,0112	0,0119	0,012	0,0146	0,0116	0,0104	0,0112	0,0113	0,0079	0,0081	0,008	0,0124	0,0113	0,0103	0,0113	0,0124	0,0148
19	0,0544	0,0502	0	0,0136	0,014	0,0149	0,018	0,0152	0,0225	0,025	0,0149	0,018	0,0203	0,0216	0,0146	0,0148	0,0211	0,0155	0,0159	0,0144
20	0,0432	0,041	0,0472	0	0,0052	0,006	0,017	0,0079	0,0139	0,0163	0,0069	0,0092	0,011	0,0114	0,0054	0,0065	0,0128	0,0066	0,007	0,0098
21	0,0441	0,0409	0,0498	0,0398	0	0,0059	0,0179	0,0079	0,0141	0,017	0,0069	0,0094	0,011	0,0125	0,0054	0,0061	0,0135	0,0067	0,0068	0,0093
22	0,0427	0,04	0,0506	0,0404	0,0381	0	0,016	0,0062	0,0134	0,0165	0,0068	0,0091	0,0109	0,0119	0,0065	0,0076	0,0145	0,0059	0,0064	0,009
23	0,047	0,046	0,0374	0,0468	0,0444	0,043	0	0,0145	0,0151	0,0172	0,0177	0,0151	0,0148	0,0148	0,0186	0,0161	0,0163	0,0153	0,0166	0,0146
24	0,0463	0,042	0,0416	0,0424	0,0404	0,0398	0,0341	0	0,0137	0,0175	0,0078	0,0095	0,0113	0,0125	0,008	0,0088	0,0141	0,0068	0,0071	0,0083
25	0,0395	0,0377	0,0539	0,0442	0,0435	0,0416	0,0468	0,0452	0	0,0087	0,0127	0,0104	0,0102	0,0097	0,0143	0,0125	0,0104	0,0119	0,0143	0,0158
26	0,041	0,0403	0,0533	0,0479	0,0467	0,0436	0,046	0,047	0,0403	0	0,0147	0,0136	0,0116	0,0109	0,0172	0,0156	0,0115	0,0156	0,0163	0,0196
27	0,0452	0,0423	0,0508	0,0414	0,0429	0,0407	0,0484	0,0417	0,0465	0,0454	0	0,0091	0,011	0,0125	0,006	0,008	0,0147	0,0074	0,0066	0,0106
28	0,0424	0,0366	0,051	0,0398	0,041	0,0402	0,0444	0,0409	0,04	0,0435	0,0432	0	0,0075	0,008	0,0099	0,0085	0,0115	0,009	0,0102	0,0119
29	0,0387	0,0365	0,051	0,04	0,0384	0,0396	0,0452	0,043	0,0391	0,0414	0,0435	0,0392	0	0,0068	0,0115	0,0098	0,011	0,0105	0,0115	0,014
30	0,0404	0,0404	0,0561	0,0442	0,043	0,0421	0,0472	0,0457	0,0414	0,0447	0,0463	0,0404	0,0384	0	0,0126	0,011	0,0087	0,0113	0,0134	0,0152
31	0,0454	0,0402	0,0482	0,0387	0,0357	0,0392	0,0455	0,0396	0,042	0,0468	0,0399	0,0402	0,0394	0,0433	0	0,0072	0,0135	0,0068	0,0071	0,0092
32	0,0398	0,039	0,0455	0,0372	0,0352	0,0378	0,0441	0,0391	0,0392	0,0403	0,0426	0,0391	0,0369	0,0395	0,0341	0	0,0113	0,0072	0,0083	0,0103
33	0,0402	0,0398	0,0529	0,0409	0,0409	0,0419	0,0449	0,0454	0,041	0,0419	0,047	0,042	0,0395	0,0432	0,0406	0,0356	0	0,0128	0,015	0,0165
34	0,0392	0,0376	0,0472	0,0384	0,0371	0,0359	0,0424	0,0394	0,0396	0,0419	0,0407	0,0386	0,0365	0,0404	0,0361	0,034	0,0372	0	0,0065	0,0079
35	0,0446	0,039	0,0473	0,0381	0,0386	0,0379	0,0435	0,0388	0,0426	0,0449	0,0369	0,0421	0,0394	0,0419	0,0394	0,0376	0,042	0,0365	0	0,0092
36	0,0489	0,0468	0,0394	0,0426	0,042	0,0411	0,0339	0,036	0,0455	0,0493	0,0438	0,0444	0,0447	0,0475	0,0423	0,0427	0,0449	0,0394	0,0426	0

Tabela S1 – ciąg dalszy: szary – dane NGS, biały – skafoldy; wartości zaokrąglone do czwartego miejsca po przecinku;

Tabela S2: Procent zmienności objaśniany przez współrzędne główne PC1 - PC4 w oddzielnych analizach dla poszczególnych terminów GOSlim.

Ontologia	Termin GOSlim	PC1	PC2	PC3	PC4
komponent	*cell wall*	0,2131	0,1454	0,1045	0,0842
	cellular component	0,4653	0,0751	0,0664	0,0508
	chloroplast	0,3890	0,0674	0,0658	0,0551
	cytoplasm	0,4369	0,0750	0,0608	0,0523
	cytoskeleton	0,6435	0,0644	0,0558	0,0376
	cytosol	0,3864	0,0787	0,0776	0,0615
	endoplasmic reticulum	0,4945	0,0896	0,0678	0,0580
	endosome	0,4568	0,0868	0,0773	0,0639
	external encapsulating structure	0,2105	0,1446	0,1032	0,0832
	extracellular region	0,2432	0,1314	0,1161	0,0631
	Golgi apparatus	0,4826	0,0832	0,0799	0,0548
	intracellular	0,5194	0,0687	0,055	0,0463
	membrane	0,4399	0,075	0,0671	0,0566
	mitochondrion	0,4178	0,0714	0,0632	0,0501
	nuclear envelope	0,7154	0,0853	0,0637	0,0497
	nucleolus	0,5148	0,1014	0,0523	0,0425
	nucleoplasm	0,6568	0,0693	0,0452	0,0351
	nucleus	0,5354	0,0685	0,0542	0,0405
	peroxisome	0,4883	0,1049	0,0653	0,0597
	plasma membrane	0,3822	0,0797	0,0726	0,0639
	plastid	0,3980	0,0689	0,0639	0,0536
	ribosome	0,2994	0,1916	0,0898	0,0479
	thylakoid	0,2029	0,1390	0,1047	0,0630
	vacuole	0,3415	0,1051	0,0654	0,0635
funkcja	*binding*	0,5246	0,0750	0,0544	0,0463
	carbohydrate binding	0,3680	0,1053	0,0856	0,0488
	catalytic activity	0,4682	0,0912	0,0672	0,0498
	chromatin binding	0,6243	0,0845	0,0667	0,0495
	DNA binding	0,4965	0,0818	0,0668	0,0482
	DNA-binding transcription factor activity	0,3062	0,1235	0,0745	0,0573
	enzyme regulator activity	0,4364	0,1008	0,0561	0,0524
	hydrolase activity	0,5258	0,0709	0,0543	0,0451

	kinase activity	0,4598	0,0944	0,0567	0,0512
	lipid binding	0,3964	0,102	0,0923	0,0671
	molecular function	0,5023	0,0809	0,0599	0,0511
	motor activity	0,8971	0,0474	0,0367	0,0315
	nuclease activity	0,6744	0,0566	0,0473	0,0346
	nucleic acid binding	0,5724	0,064	0,0594	0,0371
	nucleotide binding	0,5669	0,0639	0,0525	0,0487
	protein binding	0,5321	0,0781	0,0514	0,0425
	RNA binding	0,5951	0,0673	0,0513	0,0415
	signaling receptor activity	0,4689	0,1032	0,0755	0,0564
	signaling receptor binding	0,2735	0,1845	0,1688	0,1179
	structural molecule activity	0,3499	0,1861	0,0932	0,0458
	transcription regulator activity	0,3920	0,0968	0,0660	0,0580
	transferase activity	0,4608	0,0806	0,0690	0,0571
	translation factor activity RNA binding	0,5771	0,1191	0,0733	0,045
	translation regulator activity	0,5664	0,1206	0,0724	0,0450
	transporter activity	0,4283	0,1029	0,0786	0,0569
proces	*anatomical structure development*	0,4312	0,0841	0,0672	0,0436
	biological process	0,4774	0,0831	0,0643	0,0503
	biosynthetic process	0,5064	0,0726	0,0674	0,0482
	carbohydrate metabolic process	0,4607	0,1081	0,0695	0,0518
	catabolic process	0,4423	0,0921	0,0663	0,0547
	cell communication	0,4412	0,0781	0,0596	0,0505
	cell cycle	0,5988	0,0664	0,0532	0,0404
	cell death	0,4683	0,1986	0,1099	0,0882
	cell growth	0,4227	0,1391	0,0746	0,0714
	cell differentiation	0,4179	0,1001	0,0786	0,0716
	cellular component organization	0,5442	0,0743	0,0526	0,0469
	cellular homeostasis	0,3601	0,1229	0,0817	0,0685
	cellular process	0,4956	0,0732	0,0608	0,0507
	cellular protein modification process	0,4681	0,0774	0,0660	0,0494
	circadian rhythm	0,3738	0,2859	0,1196	0,0858
	dna metabolic process	0,6951	0,0808	0,0449	0,0295
	embryo development	0,4717	0,0966	0,082	0,0585
	flower development	0,3950	0,1126	0,0706	0,0648
	generation of precursor metabolites and energy	0,3828	0,1215	0,0898	0,0549

	growth	0,4192	0,1052	0,0734	0,0647
	lipid metabolic process	0,4399	0,0829	0,061	0,0561
	metabolic process	0,476	0,0751	0,0687	0,053
	multicellular organism development	0,4351	0,0852	0,0653	0,0432
	nucleobase-containing compound metabolic process	0,6045	0,0584	0,0492	0,041
	photosynthesis	0,2655	0,1807	0,0997	0,069
	pollination	0,3602	0,1315	0,0769	0,0661
	post-embryonic development	0,3927	0,0923	0,0654	0,0491
	protein metabolic process	0,4305	0,0769	0,0615	0,06
	regulation of gene expression, epigenetic	0,6849	0,155	0,0616	0,0453
	regulation of molecular function	0,4576	0,1058	0,0573	0,0546
	reproduction	0,3977	0,0902	0,0632	0,055
	response to abiotic stimulus	0,3562	0,1077	0,0676	0,0555
	response to biotic stimulus	0,3926	0,0943	0,0653	0,0558
	response to chemical	0,3127	0,1008	0,0775	0,0696
	response to endogenous stimulus	0,3285	0,0973	0,0734	0,0662
	response to external stimulus	0,4141	0,088	0,0585	0,0568
	response to light stimulus	0,3964	0,1086	0,0785	0,0524
	response to stress	0,4803	0,078	0,055	0,0471
	secondary metabolic process	0,2719	0,1853	0,0978	0,0809
	signal transduction	0,4322	0,0808	0,0584	0,0552
	translation	0,4217	0,1422	0,0899	0,0413
	transport	0,4689	0,0854	0,0659	0,0595
	tropism	0,3607	0,1783	0,1145	0,0883

Tabela S3: Wyniki ANOVA dla terminów GOSlim, dla poszczególnych czynników, dla PC1.

Termin GOSlim	Genotyp (G)	Dzień (D)	Temperatura (T)	G × D	G × T	D × T
komponent						
cell wall	0,7789	0,0002	**6,40e-07**	0,3393	0,2348	0,0375
cellular component	0,8989	0,343	**6,47e-05**	0,2476	0,3369	0,114
chloroplast	0,9694	0,3622	**6,63e-05**	0,271	0,5463	0,0189
cytoplasm	0,8972	0,3451	0,0001	0,2214	0,4587	0,0305
cytoskeleton	0,9981	0,1048	0,0018	0,4932	0,5758	0,0067
cytosol	0,9158	0,1904	0,0002	0,3337	0,4964	0,0289
endoplasmic reticulum	0,9131	0,1578	**3,33e-05**	0,0684	0,2524	0,0431
endosome	0,7818	0,3946	**8,93e-05**	0,0765	0,3782	0,0205
external encapsulating structure	0,806	0,0003	**5,15e-07**	0,3283	0,223	0,0328
extracellular region	0,9439	0,0002	**1,92e-11**	0,2221	0,0198	**5,91e-07**
Golgi apparatus	0,9113	0,1471	0,0003	0,1929	0,501	0,0058
intracellular	0,9115	0,5435	0,0001	0,2913	0,3104	0,0515
membrane	0,9587	0,3593	**9,18e-05**	0,231	0,3408	0,1577
mitochondrion	0,8102	0,6305	**2,80e-05**	0,2132	0,252	0,0217
nuclear envelope	0,6079	0,1641	0,0003	0,1646	0,7034	0,0796
nucleolus	0,9273	0,5308	0,0001	0,1776	0,3345	0,0984
nucleoplasm	0,9109	0,6437	0,0006	0,2533	0,3076	0,4674
nucleus	0,9312	0,6096	0,0002	0,3397	0,3949	0,1246
peroxisome	0,9671	0,6441	0,0002	0,4157	0,7218	0,3424
plasma membrane	0,9395	0,2534	**7,92e-05**	0,2468	0,3714	0,0095
plastid	0,9762	0,3725	**6,17e-05**	0,2725	0,5465	0,0215
ribosome	0,1068	0,1207	0,0025	0,8283	0,1367	0,8877
thylakoid	0,338	0,492	**3,66e-07**	0,4966	0,0388	0,3889
vacuole	0,8065	0,2385	**9,60e-05**	0,2321	0,4654	0,0714
funkcja						
binding	0,9332	0,4834	**7,19e-05**	0,2817	0,3527	0,0803
carbohydrate binding	0,7787	0,4693	**2,03e-05**	0,2248	0,4687	0,0679
catalytic activity	0,9788	0,2725	**5,46e-05**	0,2436	0,3502	0,0538
chromatin binding	0,9034	0,1951	0,0007	0,4362	0,5114	0,0344
DNA binding	0,9031	0,6299	**4,98e-05**	0,3284	0,3968	0,0464
DNA-binding transcription factor activity	0,9174	0,235	**3,57e-06**	0,1013	0,1406	0,0393
enzyme regulator activity	0,9147	0,1243	**6,65e-05**	0,1917	0,5273	0,0508

hydrolase activity	0,976	0,3892	**7,54e-05**	0,2519	0,4137	0,0565
kinase activity	0,9311	0,5407	0,0004	0,2311	0,5166	0,1126
lipid binding	0,8888	0,0883	0,0008	0,224	0,3681	0,0563
molecular function	0,9237	0,4561	**9,77e-05**	0,2724	0,3346	0,1143
motor activity	0,9949	0,1108	0,0017	0,7527	0,678	0,002
nuclease activity	0,828	0,8073	0,0002	0,3538	0,5513	0,1216
nucleic acid binding	0,9244	0,6793	0,0002	0,2736	0,4422	0,0791
nucleotide binding	0,9729	0,4474	0,0002	0,332	0,4376	0,0849
protein binding	0,9069	0,5735	0,0001	0,2825	0,3527	0,0764
RNA binding	0,8489	0,5438	0,0002	0,2062	0,5042	0,148
signaling receptor activity	0,7897	0,8877	0,0024	0,2489	0,4048	0,0792
signaling receptor binding	0,0518	0,0025	**3,85e-05**	0,482	0,3543	0,0414
structural molecule activity	0,8629	0,4134	**3,68e-05**	0,3477	0,257	0,1038
transcription regulator activity	0,9623	0,6535	**2,48e-05**	0,1665	0,1465	0,1494
transferase activity	0,9476	0,5216	0,0002	0,1904	0,3613	0,0775
translation factor activity RNA binding	0,8836	0,5643	**8,88e-05**	0,2624	0,2792	0,0761
translation regulator activity	0,8903	0,5898	**8,37e-05**	0,2669	0,2657	0,0759
transporter activity	0,9994	0,4498	**6,61e-05**	0,1633	0,2844	0,1078
proces						
anatomical structure development	0,8049	0,7754	0,0003	0,2107	0,2892	0,1024
biological process	0,9496	0,3714	**4,37e-05**	0,2209	0,3406	0,0629
biosynthetic process	0,9434	0,495	**7,01e-05**	0,2267	0,3237	0,0343
carbohydrate metabolic process	0,9583	0,1208	**1,06e-05**	0,1998	0,3038	0,0243
catabolic process	0,9591	0,2583	**6,28e-05**	0,2208	0,3517	0,1413
cell communication	0,8716	0,4224	0,0004	0,1685	0,3757	0,1519
cell cycle	0,9803	0,3834	0,0003	0,3354	0,4899	0,019
cell death	0,9944	0,1783	0,0054	0,1299	0,6914	0,1645
cell differentiation	0,9787	0,2838	0,0005	0,1241	0,4487	0,0952
cell growth	0,9934	0,2876	0,0004	0,3914	0,4345	0,173
cellular component organization	0,99	0,32	0,0005	0,3117	0,5255	0,0435
cellular homeostasis	0,7293	0,8726	0,0006	0,1322	0,1965	0,0231
cellular process	0,9519	0,4758	**5,46e-05**	0,2075	0,3472	0,0533
cellular protein modification process	0,8644	0,4913	0,0001	0,1764	0,3131	0,0907
circadian rhythm	0,7651	0,0037	**1,37e-05**	0,983	0,9453	0,0008
DNA metabolic process	0,9336	0,612	0,0003	0,4254	0,5586	0,037
embryo development	0,8472	0,7058	0,0006	0,2511	0,4714	0,0806
flower development	0,8839	0,6005	0,0004	0,1727	0,4704	0,3057

generation of precursor metabolites and energy	0,9494	0,1143	**9,36e-06**	0,2375	0,1634	0,011
growth	0,9832	0,3607	0,0002	0,3613	0,4933	0,4829
lipid metabolic process	0,9619	0,4596	**9,96e-05**	0,1772	0,3734	0,0597
metabolic process	0,9502	0,4659	**4,06e-05**	0,2457	0,3233	0,0612
multicellular organism development	0,7869	0,8994	0,0003	0,2206	0,254	0,119
nucleobase-containing compound metabolic process	0,9446	0,7933	0,0001	0,2313	0,4055	0,0721
photosynthesis	0,4778	0,0036	**8,18e-06**	0,5094	0,1051	0,001
pollination	0,7539	0,792	0,001	0,2742	0,5044	0,0531
post-embryonic development	0,7911	0,9973	0,0002	0,1962	0,3137	0,1554
protein metabolic process	0,8894	0,5354	**7,23e-05**	0,2181	0,3514	0,0782
regulation of gene expression, epigenetic	0,6734	0,2183	0,0015	0,5579	0,4056	0,0928
regulation of molecular function	0,9354	0,1687	**5,35e-05**	0,1832	0,4968	0,0381
reproduction	0,876	0,8274	0,0004	0,2648	0,3809	0,0902
response to abiotic stimulus	0,8277	0,4494	0,0001	0,2565	0,3832	0,2043
response to biotic stimulus	0,9907	0,5831	**9,23e-05**	0,2785	0,4846	0,1195
response to chemical	0,9809	0,1712	**1,94e-05**	0,313	0,2179	0,0381
response to endogenous stimulus	0,9447	0,2634	0,0001	0,3106	0,264	0,1199
response to external stimulus	0,9191	0,4121	0,0002	0,2331	0,4526	0,138
response to light stimulus	0,869	0,5769	0,0002	0,2091	0,3433	0,3289
response to stress	0,9342	0,6452	**8,86e-05**	0,3358	0,4034	0,082
secondary metabolic process	0,3232	5,25e-05	**9,10e-05**	0,4133	0,2526	0,0011
signal transduction	0,9242	0,3648	0,0003	0,1402	0,3231	0,1408
translation	0,9701	0,4895	**4,16e-05**	0,3665	0,305	0,0547
transport	0,9545	0,3462	0,0001	0,2295	0,4354	0,0699
tropism	0,7834	0,0801	0,0007	0,1073	0,7199	0,4947

Tabela S4: Wyniki ANOVA dla terminów GOSlim, dla poszczególnych czynników, dla PC2.

Termin GOSlim	Genotyp (G)	Dzień (D)	Temperatura (T)	G × D	G × T	D × T
komponent						
cell wall	0,3402	**6,86e-08**	0,0673	0,0555	0,2214	**4,43e-06**
cellular component	0,4914	**8,64e-10**	0,4779	0,0499	0,2978	**9,67e-07**
chloroplast	0,0649	0,0008	0,1539	0,8755	0,5552	0,6824
cytoplasm	0,1056	0,0011	0,9233	0,1435	0,2324	0,0003
cytoskeleton	0,5159	0,0001	0,2139	0,0473	0,2137	**6,89e-05**
cytosol	0,1439	0,0011	0,1703	0,108	0,7458	0,6534
endoplasmic reticulum	**2,11e-06**	0,0055	0,7709	0,672	0,1505	0,4134
endosome	0,2693	0,0166	0,5622	0,5098	0,1488	0,5454
external encapsulating structure	0,4824	**1,03e-07**	0,0994	0,0696	0,275	**8,50e-06**
extracellular region	0,2084	**8,96e-07**	0,0337	0,0861	0,1427	0,012
Golgi apparatus	**8,99e-05**	0,0028	0,692	0,1811	0,0384	0,84
intracellular	0,2549	0,0012	0,6534	0,2588	0,1905	**1,49e-08**
membrane	0,2488	**4,06e-08**	0,06	0,2718	0,5263	0,0102
mitochondrion	0,2094	0,1805	0,0997	0,9582	0,4239	0,0435
nuclear envelope	0,2935	0,5889	0,8941	0,082	0,7002	0,8836
nucleolus	0,7194	0,0034	0,4534	0,5099	0,6154	0,4663
nucleoplasm	0,7355	**9,93e-05**	0,551	0,2133	0,3908	0,001
nucleus	0,2017	**2,52e-05**	0,5965	0,0138	0,3711	**1,78e-09**
peroxisome	0,002	0,2465	0,1569	0,7234	0,7357	0,3669
plasma membrane	0,6056	**5,21e-07**	0,112	0,5793	0,0743	0,0197
plastid	0,1603	**7,83e-05**	0,6627	0,5686	0,1242	0,114
ribosome	0,1463	0,0171	0,0993	0,6436	0,6495	0,0125
thylakoid	0,1452	0,9082	0,8901	0,718	0,8444	0,0012
vacuole	0,798	0,0012	0,288	0,7455	0,2575	0,007
funkcja						
binding	0,0636	**4,63e-06**	0,353	0,0449	0,1549	**5,88e-09**
carbohydrate binding	0,0351	**1,80e-08**	0,7157	0,0256	0,036	**9,22e-09**
catalytic activity	0,3954	**6,59e-08**	0,688	0,1246	0,0486	**5,60e-06**
chromatin binding	0,175	0,1898	0,2016	0,0146	0,9241	0,0053
DNA binding	0,1071	0,1515	0,381	0,0176	0,1168	**2,96e-10**
DNA-binding transcription factor activity	0,673	**6,89e-09**	0,8499	0,6825	0,8425	0,0002
enzyme regulator activity	0,4229	0,0069	0,2717	0,0436	0,1652	**3,05e-07**
hydrolase activity	0,5682	**5,17e-05**	0,0462	0,084	0,0472	**2,02e-09**
kinase activity	0,0469	**9,06e-09**	0,1336	0,7982	0,1298	0,0002
lipid binding	0,2757	0,0001	0,1896	0,1115	0,9412	0,0004
molecular function	0,3424	**5,86e-10**	0,0395	0,0044	0,214	**4,03e-05**
motor activity	0,3072	0,209	0,9455	0,1059	0,025	0,2466

nuclease activity	0,494	**2,30e-05**	0,5161	0,0061	0,1145	**5,02e-05**
nucleic acid binding	0,0644	**8,68e-07**	0,1568	0,0287	0,706	**5,96e-06**
nucleotide binding	0,0003	**1,32e-09**	0,3014	0,2179	0,087	0,0097
protein binding	0,0166	**4,64e-07**	0,4604	0,0203	0,1167	**1,62e-06**
RNA binding	0,2779	0,0005	0,2418	0,45	0,4119	0,3814
signaling receptor activity	0,2274	**1,19e-05**	0,005	0,4214	0,6841	0,488
signaling receptor binding	0,6905	0,5251	0,4128	0,243	0,1995	0,288
structural molecule activity	0,0534	0,0421	0,9065	0,8418	0,5198	0,0641
transcription regulator activity	0,7425	**3,35e-10**	0,1604	0,9158	0,7193	0,0001
transferase activity	0,0053	**3,48e-14**	0,5793	0,0012	0,0045	**8,65e-05**
translation factor activity RNA binding	0,0299	0,0487	0,1696	0,9217	0,7462	0,6927
translation regulator activity	0,0223	0,0405	0,1738	0,9074	0,7762	0,688
transporter activity	0,2865	**2,64e-12**	0,0254	0,0019	0,0051	**6,02e-10**
proces						
anatomical structure development	0,7498	0,0002	0,7953	0,9279	0,665	0,4851
biological process	0,1938	**4,58e-12**	0,4473	0,0097	0,0181	**2,28e-05**
biosynthetic process	0,1573	0,2925	0,2921	0,3384	0,3138	**3,50e-08**
carbohydrate metabolic process	0,4041	**2,64e-08**	0,2602	0,0072	0,2439	0,0039
catabolic process	0,9382	**1,48e-06**	0,3309	0,0134	0,0087	**5,45e-11**
cell communication	0,0481	**3,13e-08**	0,1429	0,0435	0,0549	0,0003
cell cycle	0,0908	0,0005	0,0884	0,0065	0,1273	**1,41e-08**
cell death	0,0691	0,0357	0,079	0,1525	0,0283	0,836
cell differentiation	0,5024	0,0016	0,7295	0,0083	0,0814	**2,07e-05**
cell growth	0,5937	0,0085	0,8393	0,881	0,435	0,623
cellular component organization	0,6772	**5,96e-05**	0,6777	0,0373	0,4274	**8,14e-06**
cellular homeostasis	0,4815	**2,04e-07**	0,0392	0,6093	0,2221	0,0093
cellular process	0,2441	**9,87e-07**	0,4966	0,0412	0,1292	**4,43e-07**
cellular protein modification process	0,0414	**1,39e-06**	0,9695	0,9487	0,1523	0,0004
circadian rhythm	0,8829	0,6696	0,7952	0,1328	0,0869	0,2686
DNA metabolic process	0,0339	0,0067	0,6015	0,0139	0,5517	**6,95e-09**
embryo development	0,775	0,002	0,8607	0,6116	0,6235	0,9843
flower development	0,0601	**1,53e-07**	0,8449	0,003	0,1694	**2,69e-05**
generation of precursor metabolites and energy	0,2667	0,0432	0,3333	0,6211	0,6791	0,0218
growth	0,6663	0,0057	0,9803	0,9455	0,5645	0,9122
lipid metabolic process	0,0105	**1,61e-11**	0,9803	0,1178	0,0083	0,0294

metabolic process	0,3518	**9,97e-10**	0,4268	0,0488	0,0405	**3,25e-06**
multicellular organism development	0,7312	**5,89e-05**	0,8659	0,8433	0,5303	0,966
nucleobase-containing compound metabolic process	0,1422	**3,80e-07**	0,8466	0,0347	0,5005	**1,61e-07**
photosynthesis	0,1177	0,0518	0,0141	0,4283	0,3261	0,0003
pollination	0,4205	**2,24e-10**	0,0001	0,7787	0,0008	0,4965
post-embryonic development	0,5783	**6,70e-05**	0,5656	0,6095	0,2102	0,1286
protein metabolic process	0,0459	**1,29e-06**	0,3007	0,8726	0,636	**2,82e-05**
regulation of gene expression, epigenetic	0,016	0,165	0,231	0,4317	0,5949	0,0027
regulation of molecular function	0,3558	0,0071	0,1597	0,017	0,3223	**9,85e-08**
reproduction	0,1668	**8,31e-05**	0,8994	0,8702	0,246	0,4038
response to abiotic stimulus	0,1616	**2,35e-05**	0,074	0,1799	0,0879	**2,77e-05**
response to biotic stimulus	0,3799	0,0008	0,3022	0,4739	0,2832	0,9285
response to chemical	0,935	**2,92e-07**	0,3179	0,1166	0,2669	0,0844
response to endogenous stimulus	0,9782	**2,05e-05**	0,2411	0,233	0,7004	0,2531
response to external stimulus	0,217	0,0011	0,5606	0,7925	0,4366	0,9215
response to light stimulus	0,095	**2,90e-05**	0,1798	0,0614	0,0373	0,0003
response to stress	0,2996	**4,19e-06**	0,0883	0,1669	0,233	**9,95e-07**
secondary metabolic process	0,8388	**5,11e-05**	0,0011	0,5448	0,239	0,0103
signal transduction	0,2846	**7,84e-08**	0,0729	0,0191	0,1159	0,0001
translation	0,0453	0,0207	0,2808	0,7259	0,4444	0,1649
transport	0,1695	**5,00e-11**	0,3454	0,0079	0,0915	**6,19e-06**
tropism	0,5256	0,0899	0,4958	0,7761	0,6742	0,234

Tabela S5: Wyniki ANOVA dla terminów GOSlim, dla poszczególnych czynników, dla PC3.

Termin GOSlim	Genotyp (G)	Dzień (D)	Temperatura (T)	G × D	G × T	D × T
komponent						
cell wall	0,008	0,0013	0,0407	0,0178	0,0365	**4,83e-07**
cellular component	0,1401	0,002	0,4542	0,7514	0,4754	**1,87e-05**
chloroplast	0,3377	0,0323	0,0436	0,5015	0,1427	0,0019
cytoplasm	0,3458	0,0413	0,1694	0,8752	0,5634	0,3876
cytoskeleton	0,2816	0,1385	0,1307	0,986	0,3048	0,0159
cytosol	0,4904	0,4095	0,9566	0,4037	0,4402	0,2097
endoplasmic reticulum	0,332	0,6993	0,1043	0,7681	0,3033	0,0903
endosome	0,0068	0,0112	0,2857	0,4108	0,0009	0,0251
external encapsulating structure	0,0111	0,0009	0,0429	0,0165	0,0223	**2,39e-07**
extracellular region	0,0088	**3,08e-05**	0,0097	0,8716	0,965	**4,27e-08**
Golgi apparatus	0,2771	0,0014	0,3898	0,0365	0,3219	0,0003
intracellular	0,0873	0,0002	0,1431	0,503	0,684	0,0702
membrane	0,3859	0,0036	0,9109	0,35	0,4918	**6,33e-09**
mitochondrion	0,2	**1,57e-09**	0,5199	0,0012	0,0045	**3,37e-06**
nuclear envelope	0,2304	0,0011	0,2628	0,4808	0,0186	0,0002
nucleolus	0,2575	0,2027	0,833	0,0874	0,0775	0,2156
nucleoplasm	0,0002	0,8361	0,2315	0,1244	0,488	0,0538
nucleus	0,2773	**3,87e-06**	0,0893	0,2954	0,8535	0,0012
peroxisome	0,3581	0,0046	0,9254	0,282	0,65	0,0214
plasma membrane	0,0263	0,7124	0,0457	0,1692	0,6688	0,003
plastid	0,0957	0,6899	0,0263	0,7199	0,7161	0,0041
ribosome	0,4147	0,1293	0,001	0,944	0,9889	0,0118
thylakoid	0,7687	**4,99e-07**	0,2025	0,3906	0,7998	0,0015
vacuole	0,0003	0,687	0,7984	0,6856	0,6513	0,36
funkcja						
binding	0,2114	**1,11e-08**	0,2171	0,6232	0,2355	0,0002
carbohydrate binding	0,0233	**1,20e-06**	0,2826	0,9878	0,446	**2,33e-06**
catalytic activity	0,21	**2,19e-06**	0,2188	0,9454	0,5812	**2,11e-07**
chromatin binding	0,1113	0,1046	0,9014	0,5545	0,9323	0,533
DNA binding	0,0276	**1,56e-09**	0,052	0,1216	0,9506	0,037
DNA-binding transcription factor activity	0,8425	**1,56e-07**	0,0015	0,0192	**3,31e-05**	**4,61e-11**
enzyme regulator activity	0,1493	0,7324	0,2424	0,9954	0,3885	0,0549
hydrolase activity	0,0009	**8,78e-07**	0,0076	0,1262	0,1497	0,0194
kinase activity	0,0611	0,0003	0,7446	0,0894	0,0329	**8,24e-08**
lipid binding	0,9453	0,3748	0,023	0,8408	0,4022	0,0063
molecular function	0,4146	**2,64e-05**	0,345	0,7983	0,3713	**2,51e-07**
motor activity	0,0037	0,5195	0,5768	0,5189	0,7817	0,0611

nuclease activity	0,1056	0,0854	0,0023	0,8916	0,3937	0,3839
nucleic acid binding	0,1574	0,0017	0,0191	0,2777	0,1347	**3,74e-07**
nucleotide binding	0,9341	0,0036	0,2804	0,6969	0,8342	**3,98e-06**
protein binding	0,6257	**2,05e-05**	0,1111	0,6875	0,9741	**1,82e-05**
rna binding	0,4597	0,8559	0,1673	0,3767	0,1073	0,0064
signaling receptor activity	**7,54e-06**	0,0026	0,5898	0,6541	0,2389	0,2297
signaling receptor binding	0,0081	0,1112	0,0087	0,5707	0,0751	0,2132
structural molecule activity	0,7521	0,0678	0,2489	0,8356	0,7819	0,4454
transcription regulator activity	0,8357	0,0004	0,0065	0,3842	0,2768	**4,44e-06**
transferase activity	0,6421	0,0318	0,1359	0,9762	0,8776	**1,28e-05**
translation factor activity RNA binding	0,797	0,0656	0,9068	0,4229	0,1118	0,0007
translation regulator activity	0,8251	0,0589	0,9854	0,3971	0,1423	0,0008
transporter activity	0,0767	0,0076	0,0485	0,9694	0,1276	0,002
proces						
anatomical structure development	0,1749	0,2306	0,991	0,0059	0,0511	**2,90e-07**
biological process	0,1931	0,0012	0,8515	0,5944	0,3464	**4,16e-06**
biosynthetic process	0,0328	**2,35e-06**	0,2245	0,3937	0,889	0,9158
carbohydrate metabolic process	0,1131	0,006	0,513	0,9891	0,3497	**1,31e-05**
catabolic process	0,5537	0,0039	0,1967	0,5225	0,8303	0,1826
cell communication	0,8394	0,0045	0,023	0,6905	0,4651	0,0118
cell cycle	0,1101	0,0007	0,5341	0,8921	0,4999	0,0884
cell death	0,0141	0,0941	0,6828	0,8065	0,1695	0,1389
cell differentiation	0,5822	0,5112	0,3508	0,2174	0,4362	0,0868
cell growth	0,6954	0,3465	0,7405	0,0885	0,5519	0,0579
cellular component organization	0,1834	0,0738	0,0541	0,9705	0,4431	0,0877
cellular homeostasis	0,5359	0,183	0,708	0,21	0,5723	0,1468
cellular process	0,0909	**1,84e-05**	0,3731	0,9295	0,3875	0,0004
cellular protein modification process	0,0586	**7,38e-07**	0,3783	0,029	0,0067	**1,44e-09**
circadian rhythm	0,0356	0,5964	0,4527	0,2245	0,2778	0,7706
DNA metabolic process	0,6892	0,0061	0,0975	0,1269	0,8839	0,0605
embryo development	0,4651	0,6221	0,995	0,4667	0,2893	**4,92e-06**
flower development	0,0123	0,1064	0,604	0,9043	0,1266	0,0003
generation of precursor metabolites and energy	0,5186	**1,22e-05**	0,8619	0,2941	0,7342	0,023
growth	0,5048	0,258	0,4079	0,1145	0,4627	0,2631
lipid metabolic process	0,7813	0,6855	0,0253	0,8342	0,6516	0,6607

metabolic process	0,1666	**1,42e-05**	0,6503	0,7664	0,786	**1,98e-06**
multicellular organism development	0,0955	0,0604	0,7147	0,0034	0,0883	**1,10e-07**
nucleobase-containing compound metabolic process	0,2072	**2,07e-05**	0,2179	0,2478	0,5439	**9,46e-06**
photosynthesis	0,7901	0,0005	0,0579	0,3654	0,7377	0,928
pollination	0,1344	0,6023	0,6322	0,198	0,9293	0,2942
post-embryonic development	0,6967	0,6659	0,8746	0,1013	0,2967	**6,00e-05**
protein metabolic process	0,7314	0,0002	0,1167	0,2686	0,1272	0,0006
regulation of gene expression, epigenetic	0,2912	0,98	0,6902	0,6124	0,2007	0,1098
regulation of molecular function	0,8773	0,9825	0,0481	0,3753	0,4213	0,5164
reproduction	0,1966	0,1604	0,8717	0,0483	0,3587	**2,07e-05**
response to abiotic stimulus	0,6221	0,0821	0,4104	0,1987	0,4361	**9,08e-05**
response to biotic stimulus	0,4115	0,1342	0,0634	0,1013	0,6795	**2,59e-05**
response to chemical	0,7048	0,0473	0,0486	0,9313	0,3493	0,7914
response to endogenous stimulus	0,974	0,0066	0,0246	0,6623	0,3464	0,8298
response to external stimulus	0,0009	0,0565	0,0297	0,3831	0,1044	0,42
response to light stimulus	0,1952	0,3737	0,1512	0,0266	0,5275	0,0004
response to stress	0,2291	0,1899	0,0067	0,4015	0,9284	0,0002
secondary metabolic process	**6,56e-06**	0,3862	0,359	0,3375	0,8671	0,0153
signal transduction	0,2388	0,1373	0,1964	0,3184	0,139	0,9873
translation	0,8803	0,121	0,6674	0,7686	0,584	0,4763
transport	0,1344	0,0059	0,2529	0,6772	0,0719	**1,56e-05**
tropism	0,0208	0,6688	0,1261	0,4822	0,7258	0,0046

Tabela S6: Wyniki ANOVA dla terminów GOSlim, dla poszczególnych czynników, dla PC4.

Termin GOSlim	Genotyp (G)	Dzień (D)	Temperatura (T)	G × D	G × T	D × T
komponent						
cell wall	0,0312	0,3392	0,9122	0,9373	0,4928	0,761
cellular component	0,9608	0,0212	0,0056	0,9921	0,4765	0,5109
chloroplast	0,4415	0,1507	0,2629	0,9889	0,3253	0,001
cytoplasm	0,8355	0,5453	0,509	0,2798	0,1	0,0005
cytoskeleton	0,3626	0,8493	0,1333	0,2101	0,2177	0,7424
cytosol	0,1129	0,448	0,9006	0,1626	0,4032	**1,90e-05**
endoplasmic reticulum	0,1484	0,1353	0,6244	0,9547	0,1273	0,5377
endosome	0,2491	0,2052	0,6946	0,0349	0,3458	**9,83e-05**
external encapsulating structure	0,0405	0,3304	0,932	0,9323	0,5719	0,9676
extracellular region	0,2377	0,0849	0,8537	0,6699	0,4513	0,9174
Golgi apparatus	0,0002	0,2409	0,4053	0,8156	0,5237	0,0463
intracellular	0,8817	0,4741	0,3505	0,4593	0,4957	0,3752
membrane	0,6258	0,0064	0,0029	0,8373	0,2312	0,5251
mitochondrion	0,1442	0,4771	0,698	0,8358	0,5441	0,2342
nuclear envelope	0,3799	0,214	0,6149	0,9025	0,4687	0,6524
nucleolus	0,0285	0,227	0,876	0,309	0,6943	0,9056
nucleoplasm	0,6817	0,5062	0,6867	0,5597	0,1299	0,8709
nucleus	0,4669	0,0256	0,0167	0,3259	0,0253	0,7049
peroxisome	0,0009	0,1359	0,4026	0,9071	0,5067	0,2972
plasma membrane	0,0313	0,0056	0,6061	0,9417	0,8555	0,006
plastid	0,2629	0,0447	0,4079	0,9553	0,3239	0,0052
ribosome	0,0104	0,1821	0,4228	0,0112	0,4096	0,0053
thylakoid	0,0153	0,3502	0,4678	0,2752	0,7942	0,0199
vacuole	0,5971	0,0621	0,1497	0,2483	0,5794	0,0494
funkcja						
binding	0,7701	0,4567	0,0257	0,8762	0,4155	0,1213
carbohydrate binding	0,0002	0,0021	0,0321	0,3168	0,2731	0,4276
catalytic activity	0,7663	0,0844	0,0123	0,9216	0,4133	0,8467
chromatin binding	0,7975	0,7072	0,1081	0,048	0,945	0,0289
DNA binding	0,5572	0,0111	**8,15e-05**	0,5894	0,3294	0,1483
DNA-binding transcription factor activity	0,9357	0,5501	0,0019	0,7162	0,4583	0,054
enzyme regulator activity	0,783	0,0022	0,0311	0,3443	0,2587	0,1497
hydrolase activity	0,1301	0,0066	0,0087	0,9136	0,6366	0,1053
kinase activity	0,3313	0,7739	0,0077	0,8604	0,7507	0,5101
lipid binding	0,175	0,0142	0,8786	0,7289	0,6631	0,0063
molecular function	0,7252	0,8755	0,0572	0,6814	0,0638	0,0523
motor activity	0,6809	0,01	0,122	0,2878	0,4088	0,8195
nuclease activity	0,5336	0,2616	0,24	0,8587	0,4924	0,3169
nucleic acid binding	0,1554	0,3176	0,051	0,807	0,3837	0,9938

nucleotide binding	0,8504	0,1342	0,0108	0,477	0,7606	0,014
protein binding	0,0044	0,3311	0,3772	0,5355	0,0547	0,6513
RNA binding	0,1172	0,2323	0,4053	0,2715	0,3511	0,0006
signaling receptor activity	0,2043	0,364	0,0076	0,6124	0,0728	0,1084
signaling receptor binding	0,2538	0,3357	0,2329	0,4626	0,4222	0,5085
structural molecule activity	0,0224	0,0256	0,4217	0,0026	0,0374	**7,56e-07**
transcription regulator activity	0,3644	0,4978	0,0651	0,5198	0,1258	0,0066
transferase activity	0,2545	0,9053	0,0374	0,7782	0,8764	0,0452
translation factor activity RNA binding	0,1354	0,9305	0,3297	0,1361	0,0519	0,0023
translation regulator activity	0,2118	0,768	0,1812	0,1726	0,0371	0,0017
transporter activity	0,7919	0,1057	0,0272	0,9448	0,7491	0,0051
proces						
anatomical structure development	0,0032	0,8431	0,0034	0,1479	0,8525	0,1928
biological process	0,5265	0,535	0,0159	0,9615	0,2363	0,2513
biosynthetic process	0,942	0,0733	0,0751	0,9267	0,681	0,1301
carbohydrate metabolic process	0,7434	0,4462	0,0111	0,7101	0,2452	0,0388
catabolic process	0,4127	0,0003	0,3281	0,5086	0,8291	0,0562
cell communication	0,0041	0,2467	0,6687	0,1119	0,311	0,5588
cell cycle	0,0005	0,0826	0,7299	0,265	0,7436	0,4487
cell death	0,8202	0,9279	0,7932	0,2531	0,2644	0,1099
cell differentiation	0,0488	0,163	0,0292	0,1442	0,1287	0,5355
cell growth	0,078	0,9769	0,1631	0,2664	0,3629	0,0001
cellular component organization	0,2314	0,2023	0,7806	0,3658	0,2766	0,0439
cellular homeostasis	0,8255	0,6737	0,522	0,8927	0,8914	0,7544
cellular process	0,8361	0,034	0,0767	0,8968	0,445	0,5548
cellular protein modification process	0,4062	0,2936	0,0264	0,3493	0,459	0,7681
circadian rhythm	0,3246	0,5668	0,0537	0,4314	0,1954	0,2573
dna metabolic process	0,0045	0,4124	0,1971	0,3478	0,9869	0,0572
embryo development	0,0023	0,0042	0,8119	0,0603	0,0829	0,3941
flower development	0,0263	0,1607	0,7281	0,8416	0,2836	0,3638
generation of precursor metabolites and energy	0,3274	0,0074	0,3152	0,7329	0,4045	0,0002
growth	0,4583	0,4962	0,6945	0,1055	0,1674	0,0002
lipid metabolic process	0,9831	0,017	0,6793	0,7425	0,617	**7,88e-09**
metabolic process	0,5054	0,3562	0,0276	0,9992	0,3074	0,569
multicellular organism development	0,0067	0,4286	0,0115	0,2823	0,2478	0,4703
nucleobase-containing compound metabolic process	0,7859	0,0578	0,0302	0,999	0,1646	0,497
photosynthesis	0,8611	0,0117	0,7918	0,8096	0,6502	0,0209

pollination	0,4232	0,7338	0,826	0,7555	0,0269	0,0533
post-embryonic development	0,0002	0,1315	0,7333	0,2483	0,0651	0,2972
protein metabolic process	0,634	0,1058	0,1625	0,5036	0,4102	0,0143
regulation of gene expression, epigenetic	0,1386	0,5515	0,0151	0,6911	0,4952	0,3911
regulation of molecular function	0,2283	0,0061	0,0292	0,3392	0,051	0,0405
reproduction	**1,28e-06**	0,0385	0,6546	0,8208	0,1717	0,5087
response to abiotic stimulus	0,1076	0,0135	0,008	0,9777	0,3064	0,5948
response to biotic stimulus	0,5771	0,0545	0,6089	0,6454	0,0247	0,1735
response to chemical	0,727	0,1353	0,8515	0,501	0,9723	**7,06e-07**
response to endogenous stimulus	0,9986	0,0444	0,7974	0,2729	0,6675	**1,97e-08**
response to external stimulus	0,282	0,4074	0,7522	0,4813	0,8724	**9,10e-07**
response to light stimulus	0,1508	0,4687	0,1584	0,5236	0,2355	0,1635
response to stress	0,0362	0,0129	0,0039	0,1548	0,6145	0,9554
secondary metabolic process	0,0689	0,4366	0,0163	0,0429	0,4498	0,2234
signal transduction	0,3696	0,0027	0,1167	0,7881	0,0967	**7,24e-05**
translation	0,1595	0,1028	0,0585	0,0096	0,0633	**1,05e-05**
transport	0,8529	0,4762	0,0299	0,91	0,3245	0,4063
tropism	0,1638	0,0047	0,769	0,3229	0,8763	0,7956

LITERATURA

Alberts B., Hopkin K., Johnson A.D., Morgan D., Raff M., Roberts K., Walter P. (2019) Podstawy Biologii Komórki, część I, PWN, wydanie trzecie

Anders S., Pyl P.T., Huber W. (2015) HTSeq—a Python framework to work with high-throughput sequencing data, Bioinformatics 31 (2): 166–169, DOI: 10.1093/bioinformatics/btu638

Andrews S. (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data. Dostępne na stronie: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M.S, Sherlock G. (2000) Gene ontology: tool for the unification of biology, Nat Genet. 25 (1): 25-29, DOI: 10.1038/75556

Astbury W., Bell F. (1938) X-Ray Study of Thymonucleic Acid, Nature 141**:** 747–748, DOI: 10.1038/141747b0

Avery O.T., Macleod C.M., McCarty M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III, J Exp Med. 79 (2): 137-58, DOI: 10.1084/jem.79.2.137

Biniaz Y., Tahmasebi A., Afsharifar A., Tahmasebi A., Poczai P. (2022) Meta-Analysis of Common and Differential Transcriptomic Responses to Biotic and Abiotic Stresses in *Arabidopsis thaliana*, Plants 11: 502, DOI: 10.3390/plants11040502

Blanca M.J., Alarcón R., Arnau J., Bono R., Bendayan R. (2017) Non-normal data: Is ANOVA still a valid option?, Psicothema 29 (4): 552-557, DOI: 10.7334/psicothema2016.383

Bolger A.M., Lohse M., Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15): 2114-20, DOI: 10.1093/bioinformatics/btu170

Carpenter J.R., Smuk M. (2021) Missing data: A statistical framework for practice, Biometrical Journal 63: 915–947, DOI: 10.1002/bimj.202000196

Casamassimi A., Ciccodicola A. (2019) Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation, Int J Mol Sci. 20 (6): 1281, DOI: 10.3390/ijms20061281.

Chen S., Zhou Y., Yaru Chen Y., Gu J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (17): i884–i890, DOI: 10.1093/bioinformatics/bty560

Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szcześniak M.W., Gaffney D.J., Elo L.L., Zhang X., Mortazavi A. (2016) A survey of best practices for RNA-seq data analysis, Genome Biol 17, 13, DOI: 10.1186/s13059-016-0881-8

Dahm R. (2005) Friedrich Miescher and the discovery of DNA, Developmental Biology 278 (2): 274-288, DOI: 10.1016/j.ydbio.2004.11.028.

Deonier R.C., Tavaré S., Waterman M.S. (2005) Computational genome analysis: an introduction, Springer Science+Business Media, Inc., ISBN: 0-387-98785-1

Elandt R. (1964) Statystyka matematyczna w zastosowaniu do doświadczalnictwa rolniczego, Państwowe Wydawnictwo Naukowe

Frank H.F., Chitwood D.H. (2016) Plant chimeras: The good, the bad, and the 'Bizzaria', Developmental Biology 419 (1): 41-53, DOI: 10.1016/j.ydbio.2016.07.003.

Franklin R.E., Gosling R.G. (1953) Molecular configuration in sodium thymonucleate, Nature 171 (4356): 740-741, DOI: 10.1038/171740a0

Griffith F. (1928) The Significance of Pneumococcal Types, J Hyg (Lond) 27 (2): 113-59, DOI: 10.1017/s0022172400031879

Hershey A.D., Chase M. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage, J Gen Physiol, 36 (1): 39-56, DOI: 10.1085/jgp.36.1.39

Holley R.W., Everett G.A., Madison J.T., Zamir A. (1965) Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid, Journal of Biological Chemistry 240 (5): 2122-2128, DOI: 10.1016/S0021-9258(18)97435-1.

Love M.I., Huber W., Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biology 15: 550, DOI: 10.1186/s13059-014-0550-8.

Lu H., Gordon M.I., Amarasinghe V., Strauss S.H. (2020) Extensive transcriptome changes during seasonal leaf senescence in field-grown black cottonwood (*Populus trichocarpa* Nisqually-1), Sci Rep 10, 6581, DOI: 10.1038/s41598-020-63372-2

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung D.W., Yiu S.M., Peng S., Xiaoqian Z., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T.W., Wang J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, Gigascience 1 (1): 18, DOI: 10.1186/2047-217X-1-18

Maxam A.M., Gilbert W. (1977) A new method for sequencing DNA, Proc Natl Acad Sci USA 74 (2): 560-4, DOI: 10.1073/pnas.74.2.560

Mikołajczak K. , Kuczyńska A. , Ogrodowicz P., Kiełbowicz-Matuk A., Ćwiek-Kupczyńska H., Daszkowska-Golec A., Szarejko I., Surma M., Krajewski P. (2022) High-throughput sequencing data revealed genotype-specific changes evoked by heat stress in crown tissue of barley sdw1 near-isogenic lines, BMC Genomics 23: 177, DOI: 10.1186/s12864-022-08410-1

Morrison D.F. (1967) Multivariate statistical methods. McGraw-Hill series in probability and statistics, ISBN: 007043185X

Ondov B.D., Treangen T.J., Melsted P., Mallonee A.B., Bergman N.H., Koren S., Phillippy A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash, Genome Biol. 17 (1): 132. DOI: 10.1186/s13059-016-0997-x

Patel R.K., Jain M. (2012). NGS QC Toolkit: A toolkit for quality control of next generation sequencing data, PLoS ONE 7 (2): e30619, DOI:10.1371/ journal.pone.0030619

Patro R., Duggal G., Love M. I., Irizarry R. A., Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression, Nature Methods, 14 (4): 417–419, DOI:10.1038/nmeth.4197

Prall W., Sharma B., Gregory B.D. (2019) Transcription Is Just the Beginning of Gene Expression Regulation: The Functional Significance of RNA-Binding Proteins to Post-transcriptional Processes in Plants, Plant and Cell Physiology 60 (9): 1939–1952, DOI: 10.1093/pcp/pcz067

Sanger F., Nicklen S., Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors, Proc Natl Acad Sci USA 74 (12): 5463-7, DOI: 10.1073/pnas.74.12.5463

Schubert M., Lindgreen S., Orlando L. (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging, BMC Res Notes 9, 88, DOI: https://doi.org/10.1186/s13104-016-1900-2

Szabad J. (2021) Genetic Mosaicism, Cambridge Scholars Publishing, ISBN (13): 978-1-5275-7058-0

The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. (2023) Genetics, 224 (1): iyad031, DOI: 10.1093/genetics/iyad031

Trapnell C., Pachter L., Salzberg S.L. (2009) TopHat: discovering splice junctions with RNA-Seq, Bioinformatics, DOI:10.1093/bioinformatics/btp120

Watson J., Crick F. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid, Nature, 171, 737–738. DOI: 10.1038/171737a0

Xu Y., Jia Q., Zhou G., Zhang X., Angessa T., Broughton S., Yan G., Zhang W., Li C. (2017) Characterization of the *sdw1* semi-dwarf gene in barley, BMC Plant Biol 17, 11, DOI: 10.1186/s12870-016-0964-4

Yalamanchili H.K., Wan Y.W., Liu Z. (2017) Data Analysis Pipeline for RNA-seq Experiments: From Differential Expression to Cryptic Splicing. Curr Protoc Bioinformatics, 59: 11.15.1-11.15.21, DOI: 10.1002/cpbi.33

Zhao Ch., Sahni S. (2020) Linear space string correction algorithm using the Damerau-Levenshtein distance, BMC Bioinformatics 21 (Suppl 1): 4. DOI: 10.1186/s12859-019-3184-8

Zuur A.F., Ieno E.N., Smith G.M. (2007) Analysing Ecological Data, Statistics for Biology and Health, Springer, IBSN-13: 978-0-387-45967-7

www.NobelPrize.org a. Albrecht Kossel – Facts. Nobel Prize Outreach 2025. Widziane 31 stycznia 2025

www.NobelPrize.org b. Kary B. Mullis – Facts. Nobel Prize Outreach 2025. Widziane 31 stycznia 2025

Oświadczenie promotora pracy dyplomowej

Oświadczam, że niniejsza rozprawa doktorska została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu ws. nadania stopnia naukowego doktora nauk rolniczych w dyscyplinie rolnictwo i ogrodnictwo.

10.07.25
Data

P. Krzyżewski
Podpis promotora rozprawy

Oświadczenie autora pracy doktorskiej

Świadomy odpowiedzialności prawnej oświadczam, że rozprawa doktorska przygotowana przez mnie nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona rozprawa nie była wcześniej przedmiotem procedur związanych z uzyskaniem stopnia naukowego w wyższej uczelni lub innej uprawnionej instytucji naukowej. Oświadczam ponadto, że niniejsza wersja rozprawy jest identyczna z załączoną wersją elektroniczną.

10.07.25
Data

Maria Nuc
Podpis autora pracy