Politechnika Warszawska

WYDZIAŁ MATEMATYKI
I NAUK INFORMACYJNYCH

Praca dyplomowa magisterska

na kierunku Matematyka

w specjalności Statystyka matematyczna i analiza danych

Nowe algorytmy analizy wysokowymiarowych danych z badania metylacji

Aleksandra Brodecka

Numer albumu 280648

promotor

Dr hab. inż. prof. nzw. PW Przemysław Biecek

WARSZAWA 2017

podpis promotora

Aleksandra Pradecka

podpis autora

Streszczenie

Nowe algorytmy analizy wysokowymiarowych danych z badania metylacji.

Celem niniejszej pracy dyplomowej była analiza danych z badania metylacji. Charakterystyką takich danych jest wymiar cech - sięgający milionów sond. W pracy zostały porównane różne algorytmy mające na celu wskazanie obszarów o dużej różnicy w stopniu metylacji w dwóch próbach. Dodatkowym celem pracy było stworzenie pakietu programu R umożliwiającego kompleksową analizę danych metylacyjnych zawierającego aplikację uprzednio przeanalizowanych metod. Praca została podzielona na cztery zasadnicze części. W pierwszej z nich przedstawiono aparat matematyczny potrzebny do przeprowadzenia testów statystycznych. Kolejny rozdział zawiera omówienie problemu analizy danych z badania metylacji, przegląd istniejących technik oraz zasady aplikacji metod, które zostały zbadane w niniejszej pracy. W trzeciej części pracy znajduje się porównanie algorytmów na podstawie danych pochodzących z symulacji w oparciu o dane rzeczywiste. Ostatni rozdział zawiera opis pakietu metR, dzięki któremu możliwa jest analiza danych z dziedziny metylacji.

Słowa kluczowe: analiza danych z badania metylacji, test Wilcoxona, test t-Studenta, test Kołmogorowa-Smirnowa, regresja logistyczna, regresja logistyczna z efektami losowymi, regresja logistyczna z efektami losowymi z zadaną macierzą korelacji, pakiet statystyczny R

Abstract

New algorithms for analysis of high dimensional data from methylation studies with application to human genetics.

The aim of this work was to analyze data from methylation studies. The characteristics of such data are dimension of feaures - up to millions of DNA positions. In this work various algorithms have been compared to indentify DMR - differentially methylated region in two probes. The second additional aim was to build R package that enables comprehensive analysis of methylation data with application of previously analyzed methods. The work was divided into four main parts. The first presents the mathematical tools needed to statistical hypothesis testing. The next chapter discusses the problem of analysis data from methylation studies, an overview of the existing methods and techniques of applications processes that have been investigated in this paper. In the third part of the paper there is a comparison of algorithms based on data derived from simulation. The last section contains a description of the metR package, which enables the analysis of methylation data.

Keywords: analysis data from methylation studies, Wilcoxon test, t-Student test, Kolmogorov-Smirnov test, logistic regression, logistic regression with mixed effects, logistic regression with mixed effects with given correlation structure, statistics R software

Warszawa, dnia 30.11.2017

Oświadczenie

Oświadczam, że pracę magisterską pod tytułem „Nowe algorytmy analizy wysokowymiarowych danych z badania metylacji.", której promotorem jest dr hab. inż. prof. nzw. PW Przemysław Biecek wykonałam/em samodzielnie, co poświadczam własnoręcznym podpisem.

Aleksandra Prodecka

Spis treści

Wstęp

Analiza danych z badania metylacji to zagadnienie, o którym ostatnio słyszy się coraz częściej. Metylacja jest kluczową dla zdrowia i życia modyfikacją biochemiczną, która wpływa na działanie wszystkich układów w ludzkim organizmie. Jest ona związana z takimi procesami jak imprinting rodzicielski, inaktywacja chromosomu X, modulacja struktury chromatyny czy regulacja ekspresji genu (zob. Łukasik 2009). Celem niniejszej pracy dyplomowej była analiza danych z badania metylacji. Dotychczasowe narzędzia umożliwiające przeprowadzenie weryfikacji obszarów o zróżnicowanym poziomie metylacji zakładają istnienie dwóch grup - poddanej chorobie oraz grupie kontrolnej. Dostępne testy statyczne opierają się głównie o porównanie stopnia metylacji między wspomnianymi grupami na konkretnej pozycji chromosomu. W pracy skupiono się na podejściu dotyczącym danych pochodzących od jednego pacjenta - materiału genetycznego pobranego z miejsca poddanego chorobie oraz miejsca nie poddanego chorobie. Charakterystyką takich danych jest wymiar cech - sięgający milionów sond, które wykazują bardzo dużą wartość korelacji na sąsiednich pozycjach chromosomu. Powodem tego jest fakt, że zazwyczaj cały region obejmujący kilka tysięcy pozycji DNA ulega bądź nie metylacji. W pracy zostały porównane różne algorytmy mające na celu wskazanie obszarów o dużej różnicy stopnia metylacji bazujące na stopniu metylacji lub ilości cytozyn, które zostały zmetylowane. Dodatkowym celem pracy było stworzenie pakietu programu R umożliwiającego kompleksową analizę danych metylacyjnych zawierającego aplikację uprzednio przeanalizowanych metod.

Praca została podzielona na cztery zasadnicze części. W pierwszej z nich przedstawiono aparat matematyczny potrzebny do przeprowadzenia testów statystycznych oraz opis regresji kwantylowej na estymacji której częściowo opierał się wybór najlepszej metody podczas symulacji. Do grona przeanalizowanych metod, które są obecnie wykorzystywane w badaniach z dziedziny metylacji należą: test Wilcoxona, test t-Studenta i test Kołmogorowa-Smirnowa. Wszystkie z nich bazują jedynie na stopniu metylacji. Wprowadzono również regresję logistyczną, która nie została jeszcze wykorzystana do tego rodzaju analizy. Standardową regresję logistyczną rozszerzono również o modele mieszane, które uwzględniały bądź nie strukturę korelacji danych. Algorytmy regresyjne wykorzystują informację o ilości zmetylowanych cytozyn a wybór obszarów o zróżni-

cowanym stopniu metylacji opiera się o wartość krytyczną testu Walda dla zmiennej mówiącej o typie próby. Rozdział drugi zawiera dokładne omówienie problemu poszukiwania obszarów o różnorodnym stopniu metylacji oraz opisuje charakterystykę danych. Dostępny jest również przegląd istniejących metod oraz zasady aplikacji algorytmów, które zostały zbadane w niniejszej pracy. W kolejnej części pracy znajduje się porównanie algorytmów na podstawie danych pochodzących z symulacji. Zestawione zostały różnice stopnia metylacji oraz ich kwantyle estymowane za pomocą regresji kwantylowej jak również liczba obserwacji wśród najlepszych regionów w każdej z metod. Zostało również sprawdzone podobieństwo wyboru obszarów poprzez przeliczenie wspólnie rekomendowanych regionów dla par metod. Ostatni rozdział zawiera opis pakietu *metR*, dzięki któremu możliwa jest analiza danych z dziedziny metylacji. Zawiera on funkcje umożliwiajace wstępną obróbkę danych jak również kreację obszarów, dla których zaaplikowane zostaną metody. Metody, które są dostępne w przedstawionym pakiecie statystycznym są tożsame z metodami, dla których przeprowadzone zostały badania symulacyjne.

1. Warsztat matematyczny

1.1. Podstawowe pojęcia

W pracy zakłada się podstawową znajomość statystyki matematycznej. W celu usystematyzowania niektórych pojęć wprowadzono następujące definicje (zob.Topolski 2017 oraz Koenker 1978).
W celu testowania hipotez statystycznych wprowadzono pojęcie próby losowej oraz statystyki testowej.

Definicja 1.1. Zmienne losowe $X_1; X_2; \ldots; X_n$ nazywamy **próbą losową** rozmiaru n z rozkładu o gęstości $f(x)$ (o dystrybuancie $F(x)$) jeśli $X_1; X_2, \ \ldots; X_n$ są niezależnymi zmiennymi losowymi o wspólnym rozkładzie z gęstością $f(x)$ (z dystrybuantą $F(x)$).
Przy tak przyjętej definicji rozkład próby losowej $X_1; X_2; \ldots; X_n$ jest postaci:

$$f(x_1; x_2; \ldots; x_n) = f(x_1)f(x_2)\cdots f(x_n) = \prod_{n}^{i=1} f(x_i).$$

Definicja 1.2. Niech $X_1; X_2; \ldots; X_n$ będzie próbą losową rozmiaru n, natomiast $T(x_1; x_2; \ldots; x_n)$ funkcją przyjmującą wartości rzeczywiste lub wektorowe, której dziedzina zawiera wartości jakie może przyjąć wektor $(X_1; X_2; \ldots; X_n)$. Zmienną losową lub wektor losowy:

$$Y = T(X_1; X_2; \ldots; X_n)$$

będziemy nazywać **statystyką** a rozkład Y będziemy nazywać **rozładem statystki** Y.

Na potrzeby omówienia regresji kwantylowej wprowadzono pojęcie kwantyla zmiennej losowej oraz kwantyla próbkowego próby losowej.

Definicja 1.3. Kwantylem zmiennej losowej Y rzędu $\tau, 0 < \tau < 1$ nazywamy funkcję:

$$Q(\tau) = inf\{y : F(y) \geqslant \tau\},$$

gdzie:
F oznacza dystrybuantę Y.

Definicja 1.4. Kwantylem próbkowym próby losowej $y_1; \ldots; y_n$ rzędu $\tau, 0 < \tau < 1$ nazywamy rozwiązanie poniższego problemu:

$$\min_{\varepsilon \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \varepsilon),$$

gdzie:

F oznacza dystrybuantę Y.

1.2. Test t-Studenta dla dwóch prób

Test t-Studenta dla dwóch prób porównuje różnicę średnich w dwóch próbach losowych pochodzących z rozkładu normalnego. Wyróżnia się przypadek prób zależnych i niezależnych. W obu wariantach testu korzysta się z rozkładu t-Studenta, który jest zdefiniowany następująco:

$$T = \frac{U}{\sqrt{Z}}\sqrt{n}, \tag{1.1}$$

gdzie:

$U \sim N(0,1)$, $Z \sim \chi_n$, $U \perp\!\!\!\perp Z$ (zob. Yadolah 2008).

Niech $(X_1; X_2; \ldots; X_{n_1})$ oraz $(Y_1; Y_2; \ldots; Y_{n_2})$ oznaczają dwie niezależne próbki o rozmiarach n_1 i n_2 a μ_1 i μ_2 to średnie dla populacji, z których pochodzą próby X i Y odpowiednio.

W sytuacji dwóch niezależnych próbek statystyka testowa przybiera następującą postać:

$$T = \frac{\bar{x} - \bar{y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{1.2}$$

gdzie:

$\bar{x}$ i $\bar{y}$ to średnie wartości z próby X i Y,

$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$, gdzie S_1, S_2 oznaczają odchylenia standardowe prób X i Y.

Można pokazać, że dla hipotezy zerowej (1.2) ma w przybliżeniu rozkład t-Studenta z $n_1 + n_2 - 2$ stopniami swobody.

Możliwe są następujące hipotezy testowe:

1. $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2 \quad (T > t_{n_1+n_2-2,1-\alpha/2}$ lub $T < t_{n_1+n_2-2,\alpha/2})$

2. $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 > \mu_2 \quad (T > t_{n_1+n_2-2,1-\alpha})$

3. $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 < \mu_2 \quad (T < t_{n_1+n_2-2,\alpha})$

Sytuacja 1. sprawdza czy nie ma jakiejkolwiek statystycznej różnicy między średnimi dwóch populacji. Kolejne dwa przypadki to sprawdzenie czy jedna z populacji ma istotnie statystyczną większą średnią rozkładu niż druga. W nawiasach podano wartości, dla których hipoteza zerowa zostanie odrzucona na rzecz alternatywnej dla testu na poziomie istotności α, gdzie $t_{n,\alpha}$ oznacza kwantyl rozkładu t-Studenta o n stopniach swobody na poziomie α.

W przypadku, gdy próby X i Y są zależne porównana zostaje różnica obserwacji między próbami: $d_i = X_i - Y_i$. W tej sytuacji rozmiary prób muszą być równe. Indeksy w obu próbach najczęściej odpowiadają jednemu zjawisku/obserwacji powiązanych pewną zależnością np. pacjent przed i po podaniu leku. W tym przypadku zakłada się, że różnice obserwacji mają rozkład normalny.

W przypadku par obserwacji statystyka testowa wygląda następująco:

$$T = \frac{\bar{D} - \mu_0}{\frac{S_D}{\sqrt{n}}}, \tag{1.3}$$

gdzie:

$\bar{D}$ oznacza średnią wartość d_i,

n rozmiar próby,

S_D oznacza odchylenie standardowe obliczone dla obserwacji d_i,

μ_0 oznacza średnią wartość d_i wyszczególnioną w hipotezie zerowej.

Przy H_0 zmienna opisana w (1.3) ma rozkład t-Studenta z $n-1$ stopniami swodoby.

O rozkładzie różnic możliwe są następujące hipotezy:

1. $H_0 : \mu_0 = \mu \quad H_1 : \mu_0 \neq \mu \quad (T > t_{n-1,1-\alpha/2} \text{ lub } T < t_{n-1,\alpha/2})$
2. $H_0 : \mu_0 = \mu \quad H_1 : \mu_0 > \mu \quad (T > t_{n-1,1-\alpha})$
3. $H_0 : \mu_0 = \mu \quad H_1 : \mu_0 < \mu \quad (T < t_{n-1,\alpha})$

Sytuacja 1. sprawdza czy nie ma jakiejkolwiek statystycznej różnicy między średnią różnicą dwóch populacji a podaną wartością μ_0. Kolejne dwa przypadki to sprawdzenie czy średnia różnica wartości jest istotnie statystycznie większa lub mniejsza od μ_0. W nawiasach podano wartości, dla których hipoteza zerowa zostanie odrzucona na rzecz alternatywnej dla testu na poziomie istotności α.

1.3. Test znaków Wilcoxona

Test znaków Wilcoxona należy do grupy testów nieparametrycznych. Podobnie jak test t-Studenta dla grup zależnych, służy do porównania dwóch grup powiązanych zależnością. Stanowi

alternatywę testu t-Studenta ponieważ nie czyni założeń o rozkładzie danych.

Niech $(X_1, X_2, \cdots, X_n)$ oraz $(Y_1, Y_2, \cdots, Y_n)$ oznaczają dwie próbki o rozmiarze n. Rozważamy pary obserwacji: $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Niech $|d_i|$ oznacza różnicę absolutną między x_i a y_i:

$$|d_i| = |y_i - x_i| \quad i = 1; 2; \ldots; n \tag{1.4}$$

oraz niech $m = \sum_{i=1}^{n} \mathbf{1}\{d_i \neq 0\}$ (zob. Yadolah 2008).

Wszystkim obserwacjom $d_i \neq 0$ nadane zostają rangi od 1 do m. Najmniejsza obserwacja dostaje rangę 1, potem następne kolejno o 1 większą aż do ostatniej, która otrzymuje rangę m. W przypadku, gdy występują obserwacje, które mają takie same wartości d_i ich rangi są aktualizowane na wartość średnią rang osiąganych w grupie identycznych pomiarów.

Statystyka testowa wygląda następująco:

$$T = \frac{\sum_{i=1}^{m} R_i}{\sqrt{\sum_{i=1}^{m} R_i^2}}, \tag{1.5}$$

gdzie:

R_i oznacza rangę d_i.

Można pokazać, że zmienna T jest zmienną losową o średniej $\mu = \frac{m(m+1)}{4}$ i odchyleniu standardowym: $\sigma = \sqrt{\frac{m(m+1)(2m+1)}{24}}$. Zatem zmienna losowa $Z = \frac{T-\mu}{\sigma}$ ma w przybliżeniu rozkład normalny. Kwantyl rozkładu T może być zatem przybliżany dla odpowiednio dużych m poprzez:

$$w_\alpha = \frac{m(m+1)}{4} + z_\alpha \sqrt{\frac{m(m+1)(2m+1)}{24}},$$

gdzie:

z_α oznacza kwantyl rozkładu normalnego na poziomie α.

Zazwyczaj przybliżenia rozkładem normalnym dokonuje się dla $m > 15$. Dla mniejszych wartości m korzysta się ze specjalnie skonstruowanej tablicy Wilcoxona pozwalającej na odczytanie wartości krytycznych.

Test znaków Wilcoxona umożliwia rozważanie następujących hipotez:

1. $H_0 : d_k = 0, \quad H_1 : d_k \neq 0 \quad (T < w_{\alpha/2} \text{ lub } T > w_{1-\alpha/2})$
2. $H_0 : d_k \leqslant 0, \quad H_1 : d_k > 0 \quad (T > w_{1-\alpha})$
3. $H_0 : d_k \geqslant 0, \quad H_1 : d_k < 0 \quad (T < w_\alpha)$

Sytuacja 1. sprawdza czy nie ma jakiejkolwiek statystycznej różnicy między dwoma populacjami. Kolejne dwa przypadki to sprawdzenie czy jedna z populacji ma tendencję do osiągania mniejszych wartości niż druga. W nawiasach podano wartości, dla których hipoteza zerowa zostanie odrzucona na rzecz alternatywnej dla testu na poziomie istotności α.

1.4. Test Kołmogorowa-Smirnowa

Test Kołmogorowa-Smirnowa to test nieparametryczny służący do testowania różnicy w rozkładzie dwóch prób losowych.

Niech $(X_1; X_2; \ldots; X_n)$ i $(Y_1; Y_2; \ldots; Y_m)$ oznaczają próby pochodzące z dwóch populacji o nieznanych dystrybuantach $F(x)$ i $G(x)$, których dystrybuanty empiryczne to odpowiednio $H_1(x)$ oraz $H_2(x)$. Hipotezy testowane za pomocą omawianego testu (zob. Yadolah 2008):

1. $H_0 : F(x) = G(x)$ dla każdego x, $\quad H_1 : F(x) \neq G(x)$ przynajmniej dla jednego x
2. $H_0 : F(x) \leqslant G(x)$ dla każdego x, $\quad H_1 : F(x) > G(x)$ przynajmniej dla jednego x
3. $H_0 : F(x) \geqslant G(x)$ dla każdego x, $\quad H_1 : F(x) < G(x)$ przynajmniej dla jednego x

W pierwszym przypadku sprawdzane jest czy występuje jakakolwiek różnica w rozkładzie dwóch prób. Pozostałe przypadki odnoszą się do sprawdzenia czy któraś z prób nie ma tendencji do osiągania większych wartości niż druga. Statystyka testowa jest definiowana oddzielnie dla każdej testowanej hipotezy:

1. $T_1 = \sup_x |H_1(x) - H_2(x)|$
2. $T_2 = \sup_x [H_1(x) - H_2(x)]$
3. $T_3 = \sup_x [H_2(x) - H_1(x)]$

Hipoteza zerowa zostanie odrzucona na rzecz alternatywnej, gdy $T_1(T_2 \text{ lub } T_3) > t_{n,m,1-\alpha}$, gdzie $t_{n,m,1-\alpha}$ to wartość kwantyla z tablicy Smirnowa o parametrach $n, m, 1-\alpha$. Możliwe jest również wyliczenie wartości krytycznej bezpośrednio z rozkładu Kołmogorowa zdefiniowanego następująco:

$$K = \sup_{t \in [0,1]} |B(t)|, \tag{1.6}$$

gdzie
$B(t)$ jest mostem Browna.

Dystrybuanta rozkładu Kołmogorowa może zostać zapisana w niżej przytoczony sposób: (zob. Blackman 1956)

$$P(K \leqslant x) = \frac{\sqrt{2\pi}}{x} \sum_{i=1}^{\infty} e^{-(2i-1)^2 \pi^2 / (8x^2)}. \tag{1.7}$$

W tym wypadku hipoteza zerowa zostanie odrzucona, gdy wartość statystyki testowej $T_1(T_2 \text{ lub } T_3) > K_\alpha \sqrt{\frac{n+m}{nm}}$.

1.5. Regresja logistyczna

Model regresji logistycznej można utożsamiać z modelem liniowym dla odpowiednio przekształconej wartości średniej. Można go przedstawić następującym równaniem:

$$\pi_i = h(\beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p), \tag{1.8}$$

gdzie:

π_i to średnia wartość zmiennej odpowiedzi,

$x_1; \ldots; x_p$ to zmienne objaśniające,

$\beta_1; \ldots; \beta_p$ to współczynniki modelu,

$h(\cdot)$ to pewna funkcja przekształcająca wartości liniowe na predyktorów na wartości z przedziału $(0; 1)$.

Naturalne założenie w przypadku regresji logistycznej odnosi się do rozkładu wartości zmiennej odpowiedzi:

$$Y_i \sim b(1, \pi_i), \text{ zatem } E(Y_i) = \pi_i.$$

Natomiast dla danych zgrupowanych, czyli dla danych, w których w danej grupie mamy takie same wartości zmiennych objaśniających:

$$Y_i \sim b(n_i, \pi_i), \text{ zatem } E(Y_i/n_i) = \pi_i,$$

gdzie:

n_i to liczba obserwacji w i-tej grupie.

Najczęściej korzysta się z następującej postaci funkcji $h(\cdot)$:

$$\pi_i = h(x_i^T\beta) = \frac{\exp(x_i^T\beta)}{1 + \exp(x_i^T\beta)}. \tag{1.9}$$

Na potrzeby modelu została wprowadzona funkcja logitowa, charakteryzująca się następującą zależnością : $logit(x) = \log \frac{x}{1-x}$. Można szybko wykazać, że $logit(\pi_i) = x_i^T\beta$.

Estymacja współczynników w modelu regresji logistycznej odbywa się najczęściej za pomocą maksymalizacji funkcji wiarygodności. W przypadku danych zgrupowanych Y_i wskazuje na liczbę sukcesów a nie, czy dane zdarzenie wystąpiło. Niech n_i oznacza liczbę obserwacji, które mają i-ty poziom zmiennych objaśniających. Wtedy $Y_{i1}, \ldots, Y_{in_i}$ to zmienne binarne mówiące o wystąpieniu danego zdarzenia w przypadku pojedynczej obserwacji. Mają one rozkład $bin(1, \pi_i)$.

Przy założeniu o niezależności obserwacji można otrzymać następującą własność, która tłumaczy postać rozkładu Y_i w sytuacji danych zgrupowanych:

$$\forall j = 1, \ldots, n_i \quad Y_{ij} \sim bin(1, \pi_i) \Rightarrow Y_i = Y_{1i} + \ldots + Y_{n_i i} \sim bin(n_i, \pi_i).$$

W modelu indywidualnym funkcja wiarygodności przyjmuje następującą postać:

$$L(\beta) = \prod_{i=1}^{N} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, \tag{1.10}$$

gdzie:

N to liczba obserwacji.

W modelu zgrupowanym można wykorzystać informację o liczbie poziomów zmiennych objaśniających (K):

$$L(\beta) = \prod_{i=1}^{N} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \prod_{i=1}^{K} \prod_{j=1}^{n_i} \pi_i^{y_{ij}} (1 - \pi_i)^{1-y_{ij}} = \prod_{i=1}^{K} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}. \tag{1.11}$$

W przypadku skupienia się na liczbie sukcesów w modelu zgrupowanym funkcje wiarygodności będą różnić się nieistotnie, czynnikiem $\binom{n_i}{y_i}$, który jest niezależny od β:

$$\tilde{L}(\beta) = \prod_{i=1}^{K} \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}. \tag{1.12}$$

Korzystając z postaci funkcji wiarygodności (1.10) jej logarytm możne zostać zapisany jako:

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{N} y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i). \tag{1.13}$$

Zlogarytmowaną funkcję wiarygodności można zapisać również za pomocą wektora $\underline{\beta}$ na podstawie następujących własności:

$$\log \pi_i = x_i^T \beta - \log(1 + \exp(x_i \beta)), \tag{1.14}$$

$$x_i^T \beta = logit(\pi_i) = \log(\pi_i) - \log(1 - \pi_i) \tag{1.15}$$

I podstawiając $\log(1 - \pi_i) = -\log(1 + \exp(x_i^T \beta))$ do (1.13) można otrzymać:

$$l(\underline{\beta}) = \sum_{i=1}^{n} y_i x_i^T \beta - \log(1 + x_i^T \beta). \tag{1.16}$$

Zatem funkcja score ma następującą postać:

$$U(\beta) = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - \pi_i) = X' y - x' \pi, \tag{1.17}$$

gdzie:

$\pi' = (\pi_1, \ldots, \pi_N)'$.

Dla q-elementu wektora β zaś:

$$\frac{\partial l}{\partial \beta_q} = \sum_{i=1}^{N} x_{iq}(y_i - \pi_i) \tag{1.18}$$

oraz kolejno :

$$\frac{\partial l}{\partial \beta_q \beta_{q'}} = -\sum_{i=1}^{N} x_{iq} x_{iq'} \pi_i (1 - \pi_i). \tag{1.19}$$

Powyższy wynik można uogólnić do postaci macierzowej:

$$\frac{\partial l}{\partial \beta \partial \beta} = -X'WX, \tag{1.20}$$

gdzie:
$W = diag(\pi_i(1 - \pi_i))$.

Wyprowadzenia $(1.13 - 1.20)$ są publikowane w wielu artykułach na temat regresji logsitycznej (m.in. zob. Czepiel 2015).
Inaczej niż w przypadku regresji liniowej, w celu estymacji $\underline{\beta}$ za pomocą maksymalizacji funkcji wiarygodności konieczne jest rozwiązywanie równań nieliniowych ze względu na postać otrzymaną w (1.17). Najczęstszą metodą rozwiązywania równań nieliniowych jest algorytm Newtona-Rapsona. Algorytm ten polega na iteracyjnym poszukiwaniu miejsca zerowego funkcji. Idea tego algorytmu jest następująca:

$$\frac{df}{dx} x^{(n-1)} = f'(x^{(n-1)}) \approx \frac{f(x^{(n)}) - f(x^{(n-1)})}{x^{(n)} - x^{(n-1)}}.$$

Jeśli $x^{(n)}$ (n-ta iteracja) to miejsce zerowe f, to $f(x^{(n)}) = 0$.

Algorytm Fisher-Scoring jest również wykorzystywany w celu znalezienia minimum funkcji. Korzysta on jednak z wartości oczekiwanych macierzy informacji a nie tak jak algorytm N-R - wartości obserwowanych macierzy informacji. W przypadku modelu logistycznego obie macierze są równoważne, zatem podane algorytmy również.
Dokładny mechanizm działania algorytmu N-R jest następujący:
Niech $u(\underline{\beta}) = \left(\frac{\partial l(\beta)}{\partial \beta_1}; \ldots; \frac{\partial l(\beta)}{\partial \beta_p}\right)'$, $H(\beta) = [h_{q,q'}]$, gdzie $h_{q,q'} = \frac{\partial^2 l(\beta)}{\partial \beta_q \partial \beta_{q'}}$. W kroku t następuje aproksymacja $l(\beta)$ poprzez wyrazy rozwinięcia Taylora do rzędu 2.

$$l(\underline{\beta}) \approx l(\underline{\beta}^{(t)}) + u^{(t)T}(\underline{\beta} - \underline{\beta}^{(t)}) + \frac{1}{2}(\underline{\beta} - \underline{\beta}^{(t)})^T H^{(t)} (\underline{\beta} - \underline{\beta}^{(t)}). \tag{1.21}$$

$$\frac{\partial l}{\partial \beta} \approx u^{(t)} + H^{(t)}(\underline{\beta} - \underline{\beta}^t) = 0. \tag{1.22}$$

zatem, przy założeniu, że macierz $H^{(t)}$ jest nieosobliwa:

$$\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1}u^{(t)}. \tag{1.23}$$

Podstawiając zależności: (1.17), (1.20) otrzymuje się ostatecznie:

$$\beta^{(t+1)} = \beta^{(t)} + (X'WX)^{-1}|_{\beta=\beta^{(t)}} \cdot X'(y-\pi)|_{\beta=\beta^{(t)}} = (X'WX)^{-1}X'W[X\beta + W^{-1}(y-\pi)]. \tag{1.24}$$

Powyższa postać to postać ważonej iteracyjnej metody najmniejszych kwadratów zastosowana dla wag W i zmiennych objaśniających $z = X\beta^{(t)} + W^{-1}(z-\pi)$.

Uwaga 1.5. Gdy $n \to \infty$ w przypadku danych zgrupowanych $\hat{\beta}$ ma rozkład asymptotycznie normalny.

1.6. Regresja logistyczna z efektami losowymi

Standardowy model logistyczny, który nie uwzględnia efektów losowych można zapisać następująco:

$$logit(\pi_i) = x_i'\beta. \tag{1.25}$$

Natomiast model logistyczny uwzględniający efekt losowy:

$$logit(\pi_i) = x_i'\beta + z_i'u. \tag{1.26}$$

W tym przypadku x_i to nadal wartości zmiennych obserwowanych a β to wielkości efektu dla zmiennych obserwowanych. Nowy czynnik występujące we wzorze to $z_i'u$.
u to efekt losowy zmiennej, którego wartości nie są estymowane bezpośrednio. Wyliczone natomiast zostają parametry rozkładu. Najczęściej zakłada się, że $u \sim N(0, \sigma^2 D(\theta))$, gdzie θ to nieznany wektor $c \times 1$. W powyższym wzorze z_i oznacza wektor zmiennych objaśniających dla efektów losowych.

Rozkład warunkowy y_i w logistycznym modelu mieszanym wygląda następująco:

$$Y|u \sim B(1, x'\beta + z'u). \tag{1.27}$$

Ostatecznie:

$$\pi_i = E(y_i|u) = \frac{\exp(x_i^T\beta + z_i^T u)}{1 + \exp(x_i^T\beta + z_i^T u)} \text{ oraz } Var(y_i|u) = \frac{\phi}{a_i}V(\mu_i), \tag{1.28}$$

gdzie:

ϕ to parametr rozproszenia,

a_i to waga a priori,

$V(\cdot)$ to funkcja wariancji.

Żeby otrzymać rozkład brzegowy Y konieczne jest całkowanie gęstości warunkowej względem u na podstawie poniższej zależności:

$$f_y = \int f_{y|u} f_u \quad du.$$

Zatem:

$$f_y = \int \pi_i^{y_i} (1-\pi)^{1-y_i} \frac{1}{(2\pi)^q |G|} \exp(-\frac{1}{2} u' G^{-1} u) du, \tag{1.29}$$

gdzie:

$\pi_i = x_i'\beta + z_i'u$,

$G = \sigma^2 D(\theta)$,

q oznacza długość wektora u.

Bardzo często powyższej całki nie można wyliczyć wprost, dlatego stosuje się różnego rodzaju przybliżenia w celu późniejszej estymacji parametrów. Można również rozważać zamiast marginalnej funkcji wiarygodności tzw. pseudo funkcję wiarygodności zapisaną jako:

$$L = \frac{1}{\sqrt{(2\pi)^q |G|}} \int \exp[-\frac{1}{2\phi} \sum_{i=1}^{N} d_i(y_i \mu_i) - \frac{1}{2} u^T G^{-1} u] \quad du, \tag{1.30}$$

gdzie:

$d_i(y, \mu) = -2a_i \int_y^u \frac{y-u}{V(u)} du$ oznacza dewiancję jednostkową.

Powyższego wyrażenia również nie można wyliczyć wprost, ale istnieje przybliżenie za pomocą aproksymacji Laplace'a, które prowadzi do prostej formy:

$$PQL(\beta, u) = -\frac{1}{2\phi} \sum_{i=1}^{N} d_i(y_i, \mu_i) - \frac{1}{2} u^T G^{-1} u. \tag{1.31}$$

Wyżej przedstawione równanie opisuje penalizowaną funkcję wiarygodności. Została ona uzyskana poprzez aproksymację Laplace'a (1.29) wokół warunkowej mody korzystając z rozwinięcia Taylora wokół $u = 0$. PQL to najszybsza, najbardziej elastyczna ale również najmniej dokładna metoda stosowana w estymacji efektów mieszanych. Nie ma ograniczeń ze względu na liczbę efektów losowych. Ma jednak dwie zasadnicze wady. Ze względu na małą dokładność daje głównie obciążone estymatory wariancji, szczególnie gdy liczba informacji per próbka jest niewielka np. w przypadku danych zgrupowanych w regresji logistycznej. PQL to jedynie pewne przybliżenie funkcji wiarygodności. Nie będzie zatem właściwe korzystanie z testu ilorazu wiarygodności oraz w zależności od oprogramowania, wnioskowanie na podstawie testu Walda.

Możliwe jest również zastąpienie PQL poprzez kwadratowe rozwinięcie całki z (1.29) w punkach (zob. Breslow, Clayton 1993):

$$\hat{u} = \arg\min PQL(\beta, u) \text{ dla ustalonego } \beta,$$
$$\hat{\beta} = \arg\min PQL(\beta, u) \text{ dla ustalonego } u.$$

Mając wartości nieznanych parametrów u i β można przejść do estymacji θ za pomocą metody REML. Estymatory REML mają zazwyczaj mniejsze obciążenie niż estymator ML w modelach mieszanych. W najprostszej formie w pierwszym kroku estymowane są efekty losowe przez rozważanie takich kombinacji liniowych danych, które usuwają efekty stałe poprzez zrzutowanie obserwacji na podprzestrzeń $X^{\perp}$. Następnie dla danych po zrzutowaniu maksymalizuje się funkcję wiarygodności na ograniczonej przestrzeni estymując efekty losowe. W ostatnim kroku wraca się do pierwotnego zagadnienia i estymuje się efekty stałe.

Po rozwinięciu całki z (1.29) następuje aproksymacja θ na podstawie REML z wektorem:

$$y_i^* = g(y_i) = g(\hat{u}_i) + (y_i - \hat{\mu}_i)g'(\hat{\mu}_i), \tag{1.32}$$

gdzie:
$\hat{u}_i = h(x_i^T\hat{\beta} + z_i^T\hat{u})$
a y_i można zapisać w postaci modelu liniowego:

$$y_i^* = X_i\hat{\beta} + Z_i\hat{u} + \epsilon, \tag{1.33}$$

gdzie:
$\epsilon_i \sim N(0, W_i^{-1})$,
W_i jest macierzą diagonalną z elementami: $w_i = \{V(\hat{\mu}_i)(g'(\hat{\mu}_i))^2\}$.

Dokładna procedura polega na:

1. Dla zadanych θ i u, estymowany zostaje efekt stały przez rozwiązanie poniższych równań:

$$\sum_{i=1}^{N} X_i^T V_i^{-1} X_i \beta = \sum_{i=1}^{N} X_i^T V_i^{-1} y_i^*,$$

   gdzie:
   $V_i = W_i^{-1} + Z_i G Z_i^T$.

2. Efekty losowe są estymowane przez:

$$\hat{u} = \sum_{i=1}^{N} G Z_i^T V_i^{-1}(y_i^* - X_i\hat{\beta}).$$

3. Estymator REML jest dany przez:

$$\hat{\theta}_s = \frac{\sum_{n\in Q_s} \hat{u_n^2}}{\sum_{n\in Q_s}(1 - t_{nn})} \text{ dla } s = 1; \ldots; c,$$

gdzie:

$Q_s = \{n : \sum_{i=1}^{s-1} q_i < n \leqslant \sum_{i=1}^{s} q_i\}$,

$S = W - WX(X^TWX)^{-1}X^TW$,

t_{nn} jest n-tym elementem na diagonali macierz $T = (I - Z^TSZG)^{-1}$.

4. Wracamy do kroku 1 aż do uzyskania zbieżności algorytmu.

Ostatecznie, możliwe jest również obliczenie macierzy kowariancji estymatorów w punktach: $\alpha = \hat{\alpha}$ i $\beta = \hat{\beta}$.

$Cov(\hat{\beta}) = \{\sum_{i=1}^{N} X_i^T V_i^{-1} X_i\}$, $Cov(\hat{\theta}) = H^{-1}$. Macierz H składa się z następujących elementów:

$$h_{st} = \frac{1}{2} \sum_{i \in Q_s} \sum_{j \in Q_t} (Z_{(i)}^T P Z_{(j)})^2,$$

gdzie:

$Z_{(i)}$ jest i-tym wierszem Z,

$P = V^{-1} - V^{-1}XCov(\hat{\beta})X^TV^{-1}$,

q_i to liczba poziomów i-tego efektu losowego.

Szczegółowy opis algorytmu znajduje się w (Harville 1977 oraz Jang 2006).

Innym rozwiązaniem w estymacji GLMM jest posłużenie się algorytmami Monte Carlo, które korzystają z łańcuchów Markowa. Opierają się one o podejście Bayesowskie, które stara się estymować rozkład a posteriori parametrów niż estymatory największej wiarygodności jak i samą funkcję wiarygodności. Algorytm działa trochę wolniej niż algorytmy przedstawione powyżej i wymaga podania większej ilości parametrów do optymalizacji ale już po pojedynczym uruchomieniu algorytmu mamy informację o wartościach średnich, medianach i przedziale ufności poszukiwanych parametrów. Innym możliwym sposobem posługującym się algorytmami Monte Carlo jest metoda EM (zob. Chen 2012). Algorytm EM traktuje efekty losowe jako brakujące dane a zmienne odpowiedzi jako obserwowane, (y, u) stanowią zaś pełną informację. W części E (r+1)-iteracji liczona jest wartość oczekiwana:

$$Q(\xi|\xi^{(r)}) = E(log f(y, u, \xi)|y, \xi^{(r)}) = \int \log f(y, u, \xi) f(u|y; \xi^{(r)}) du \tag{1.34}$$

Natomiast w następnym następuje maksymalizacja (1.34) ze względu na wektor parametrów $\xi = (\beta, \theta, \phi)$. Zazwyczaj (1.34) nie da się obliczyć wprost, dlatego następuje jego przybliżenie. Jeżeli jest możliwe otrzymanie losowej próbki $(u^{(1)}, u^{(2)}, \ldots, u^{(L)})^T$ z $f(u|y; \xi^{(r)})$ otrzymuje się je poprzez:

$$Q_L(\xi|\xi^{(r)}) = \frac{1}{L} \sum_{l=1}^{L} \log f(y, u^{(l)}, \xi) \tag{1.35}$$

Iteracja algorytmu EM kończy się, gdy w momencie osiągnięcia zbieżności i maksymalizacji marginalnej funkcji wiarygodności.

1.7. Regresja kwantylowa

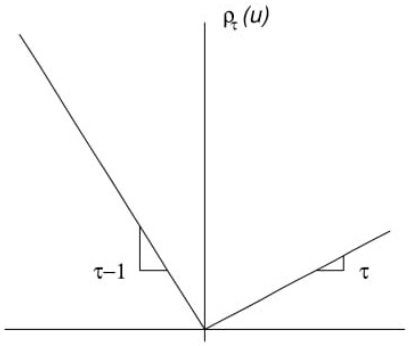
Za pomocą regresji kwantylowej możliwe jest oszacowanie różnych funkcji kwantylowych warunkowej dystrybuanty zmiennej objaśnianej. Wprowadznie różnych kwantyli regresji przewiduje pełniejszy opis podstawowych rozkładów warunkowych niż wykorzystanie estymatora najmniejszych odchyleń bezwzględnych, który umożliwia jedynie opis warunkowej mediany.

Podstawową zaletą posłużenia się regresją kwantylową przy estymacji warunkowego rozkładu jest jej odporność na obserwacje odstające. Ta metoda analizy jest szczególnie przydatna, gdy warunkowy rozkład jest asymetryczny, ucięty czy też gdy nie spełniania założenia o normalności.

Regresja kwantylowa opiera się na wyznaczeniu kwantyla odpowiedniego rzędu. Określenie kwantyla rzędu τ można sformułować jako problem optymalizacyjny. Dla każdego $0 < \tau < 1$ należy zdefiniować następującą funkcję:

$$\rho_\tau = u(\tau - I(u < 0)), \tag{1.36}$$

która została zilustrowana na rysunku 1.1.



Rysunek 1.1: Ilustracja wyznaczania kwantyla jako problemu optymalizacyjnego.

Minimalizacja wartości oczekiwanej wyrażenia $\rho_\tau(Y - \varepsilon)$ ze względu na $\varepsilon(\tau)$ prowadzi do funkcji $Q(\tau)$ zapisanej w definicji 1.4.

Idea regresji kwantylowej została zaczerpnięta z estymacji bezwarunkowej średniej, przez minimalizację następującego wyrażenia:

$$\hat{\mu} = \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} \sum (y_i - \mu)^2. \tag{1.37}$$

Powyższa estymacja może być rozszerzona do liniowej funkcji warunkowej średniej $E(Y|X = x) = x'\beta$ przez rozwiązanie:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum (y_i - x_i'\beta)^2. \tag{1.38}$$

Natomiast liniowa warunkowa funkcja kwantylowa $Q_Y(\tau|X = x) = x_i'\beta(\tau)$ może być obliczona zgodnie z poniższym:

$$\hat{\beta}(\tau) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum \rho_\tau (y_i - x_i'\beta)^2. \tag{1.39}$$

Przy regresji warunkowej mediany minimalizowana zostaje zatem następująca funkcja celu:

$$0.5 \sum (y_i - x_i'\beta)^2.$$

Dla dowolego kwantyla rzędu τ minimalizacji podlega zaś:

$$\sum_{i:y_i \geqslant x_i\beta} \tau (y_i - x_i'\beta)^2 + \sum_{i:y_i \leqslant x_i\beta} (1-\tau)(y_i - x_i'\beta)^2.$$

Kluczową własnością regresji kwantylowej jest to, że dla każdej funkcji monotonicznej $h(\cdot)$ zachodzi:

$$Q_{h(T)}(\tau|x) = h(Q_T(\tau|x)). \tag{1.40}$$

Wynika to następującej obserwacji:

$$P(T < t|x) = P(h(T) < h(t)|x). \tag{1.41}$$

Własność (1.41) nie zachodzi w przypadku estymacji metodą najmniejszych kwadratów, ponieważ w ogólności $E(h(T)|x) \neq h(E(T|x))$.

Następną równie ważną własnością jest odporność na występowanie obserwacji odstających. Wynika to z tego, że oszacowanie funkcji kwantylowej zależy wyłącznie od wartości zmiennych położonych blisko wskazanego kwantyla. Niech $\hat{\beta}(\tau)$ oznacza rozwiązanie bazujące na obserwacjach $\{y, X\}$. Jedynie zmiana znaku reszt wpłynie na oszacowanie współczynników regresji. Nie ma natomiast znaczenia wartość bezwzględna reszt z modelu.

Asymptotyczne zachowanie procesu regresji kwantylowej $\{(\hat{\beta})(\tau) : \tau \in (0,1)\}$ zostało opisane w twierdzeniu 1.6.

Twierdzenie 1.6. Niech $\{\beta_T^*(\theta_1); \ldots; \beta_T^*(\theta_M)\}$ z $0< \theta_1 < \theta_2 < \ldots < \theta_M < 1$ oszacowania pochodzące z regresji kwantylowej. Niech $\xi(\theta) = F^{-1}(\theta), \boldsymbol{\xi}(\theta) = (\xi(\theta); 0; \ldots; 0) \in \mathbb{R}^k$ oraz $\boldsymbol{\xi}_T^*(\theta) = \beta_T^*(\theta) - \beta$. Załóżmy, że:

1. F jest ciągła i ma ciągłą i dodatnią gęstość f w $\xi(\theta_i), i = 1; \ldots; M$ oraz

2. $x_{1t} = 1 : t = 1, 2, \ldots$ oraz $\lim_{T\to\infty} T^{-1}X'X = Q$ jest dodatnio określoną macierzą.

Wtedy:

$$\sqrt{T}[\xi_T^*(\theta_1) - \xi(\theta_1), \ldots, \xi_T^*(\theta_M) - \xi(\theta_M)]$$

zbiega do MK-wymiarowego rozkładu normalnego o średniej 0 i macierzy kowariancji $\Omega(\theta_1, \cdots, \theta_M; F) \otimes Q^{-1}$, gdzie Ω jest macierzą kowariancji kwantyli próby losowej M-wymiarowej o dystrybuancie F.

Szczególnym przypadkiem regresji kwantylowej jest regresja medianowa. Bez straty ogólności można założyć, że $F(0) = 1/2$, zatem $\xi(1/2) = 0$. Wtedy rozkład asymptotyczny zmiennej $\sqrt{T}(\beta*(1/2) - \beta)$ jest K-wymiarowym rozkładem normalnym ze średnią 0 i macierzą kowariancji $[2f(0)]^{-2}Q^{-1}$.

2. Postać problemu i metody

2.1. Przedstawienie problemu

Problem poruszony w pracy dotyczy analizy wysokowymiarowych danych z badania metylacji. Metylacja jest poreplikacyjną enzymatyczną modyfikacją DNA, której najczęstszym produktem jest 5-metylocytozyna. Metylacja jest niezwykle istotnym mechanizmem, ponieważ jest związana z takimi procesami jak imprinting rodzicielski, inaktywacja chromosomu X w komórkach samic ssaków łożyskowych, regulacją ekspresji genu czy też z modulacją struktury chromatyny (zob. Łukasik 2009). Poziom metylacji jest również wykorzystywany do przewidywania długości życia.

Najpopularniejsza metodą detekcji metylacji to chemiczna modyfikacja DNA wodosiarczynem sodu:

- niezmetylowana cytozyna przekształcana jest do urazylu,
- zmetylowana cytozyna nie zostaje przekształcona.

Następnie fragmenty DNA są poddawane allelospecyficznej rekacji PCR (reakcja łańcuchowej polimerazy) lub sekwencjonowaniu:

- niezmetylowana cytozyna pojawia się jako tymina,
- zmetylowana cytozyna pojawia się jako cytozyna.

W wyniku powyższych operacji otrzymuje się ostatecznie następującą postać danych:

chromosom	pozycja	meth	unmeth
chr5	1056	49	1
chr14	199421	22	24
chr16	4526	8	1
chr21	10993	38	12

Tablica 2.1: Przykładowa postać danych

Kolejne kolumny oznaczają chromosom, pozycję chromosmu oraz liczbę cytozyn, które na danej pozycji i chromosomie uległy metylacji (meth) oraz liczbie cytozyn, które na danej pozycji

nie zostały zmetylowane (unmeth). Przykładowo, pierwszy wiersz zamieszczonej tabeli wskazuje, że na pozycji 1056. w chromosomie piątym cytozyna uległa metylacji 49 razy, natomiast nie została zmetylowana jednokrotnie. Suma dwóch podanych wartości, czyli w tym przypadku 50 mówi o liczbie wszystkich odczytów reakcji w danym miejscu DNA.

Warto zwrócić uwagę na dwie rzeczy dotyczące charakteru danych. Pierwsza z nich dotyczy pozycji chromosomu dla których zebrana została informacja o przeprowadzeniu metylacji. Proces metylacji przebiega w cytozynie a więc nie we wszystkich z zasad pirymidynowych występujących w DNA (adenina, guanina, tymina). Z tego względu, oraz z powodu pewnych ograniczeń pomiarowych nie dla każdej pozycji chromosomu dostępne będą dane o procesie metylacji. Pomimo braku pomiarów na niektórych pozycjach, wymiar otrzymanej próbki jest dosyć spory. Przykładowe próbki, dla których przeprowadzono analizy mają odpowiednio 8 i 11 mln obserwacji.

Druga uwaga dotyczy procesu sekwencjonowania wodosiarczynem sodu. Nie jest możliwa jednakowa ilość odczytów reakcji dla poszczególnych pozycji. Najczęściej możliwe jest przeprowadzenie jedynie kilku reakcji wraz z poprawnym odczytem. Mediana odczytów z przykładowych danych wynosi 2, natomiast maksymalne wartości przekraczają kilka tysięcy.

Zasadniczym celem pracy jest opracowanie i weryfikacja algorytmów służących do identyfikacji obszarów, dla których proces metylacji różni się w zależności od próby. Przykładowe dane dotyczą przypadku informacji pochodzących od jednego pacjenta, gdzie próby mają odmienne miejsce pobrania materiału genetycznego. Mogą być to na przykład miejsca zmienione chorobowo i zdrowe tkanki. Porównanie takich prób będzie pozwalało odpowiedzieć na pytanie, które regiony DNA są odpowiedzialne za daną postać choroby. W pracy skupiono się na analizie obszarów, których długość wynosi zazwyczaj 1000bp. Oznacza to, że minimum i maksimum pozycji genomu analizowanego obszaru nie przekracza 1000 jednostek. Rozważanie tak krótkich obszarów w stosunku do długości całego chromosomu wynika z dużej korelacji między próbami. Tabela 2.2 przedstawia korelację wartości $\frac{meth+1}{unmeth+1}$ między dwiema próbkami na tych samych pozycjach w rozróżnieniu na chromosomy.

chr	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12	chr13
korelacja	0,939	0,939	0,947	0,939	0,940	0,945	0,933	0,935	0,933	0,934	0,936	0,939	0,936
chr	chr14	chr15	chr16	chr17	chr18	chr19	chr20	chr21	chr22	chrM	chrX	chrY	
korelacja	0,941	0,943	0,929	0,932	0,932	0,934	0,931	0,919	0,923	0,809	0,918	0,937	

Tablica 2.2: Korelacja danych z badania metylacji w podziale na chromosomy

Wszystkie wartości korelacji pomijając chromosom M są powyżej 0,918. Trudno więc spodziewać się obszarów o dużej długości, gdzie poziom metylacji będzie się zasadniczo różnił. Poza tym,

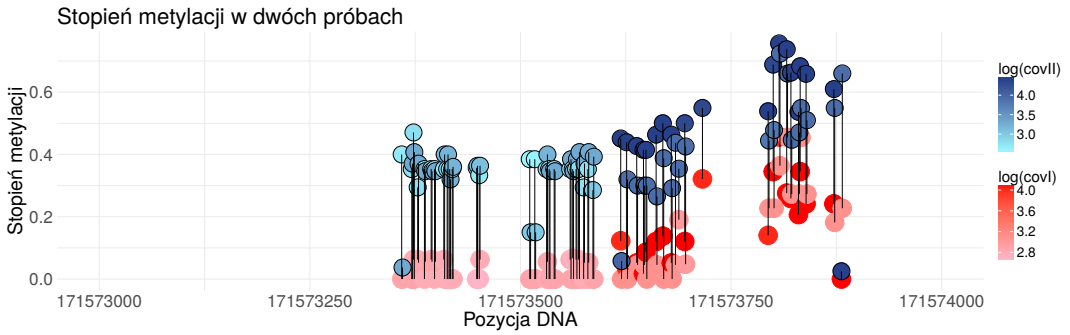
jeżeli różnica metylacji występuje na przykład na początku genu, tzw. promotorze można już wstępnie asocjować występowanie choroby z nieprawidłowym funkcjonowaniem danego genomu. Jak wiadomo, długości genów sięgają nawet 1mln, ciężko więc byłoby wykryć różnicę metylacji na tak długim regionie.

W pracy posługiwano się dwoma pojęciami: stopniem metylacji, który został zdefiniowany przez:

$$meth.rate = \frac{meth}{meth + unmeth} \tag{2.1}$$

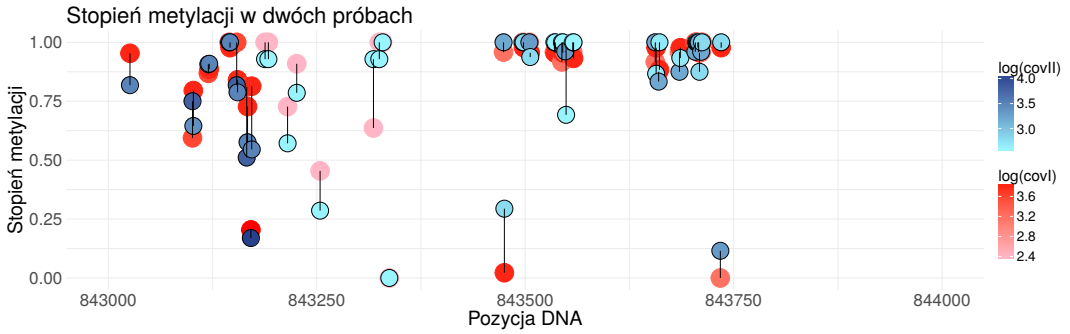
oraz stopniem pokrycia, który mówi o liczbie pozycji, dla których zgromadzono dane w analizowanym obszarze.

Rysunek 2.1 przedstawia przykład obszaru o zróżnicowanym stopniu metylacji w dwóch próbach.



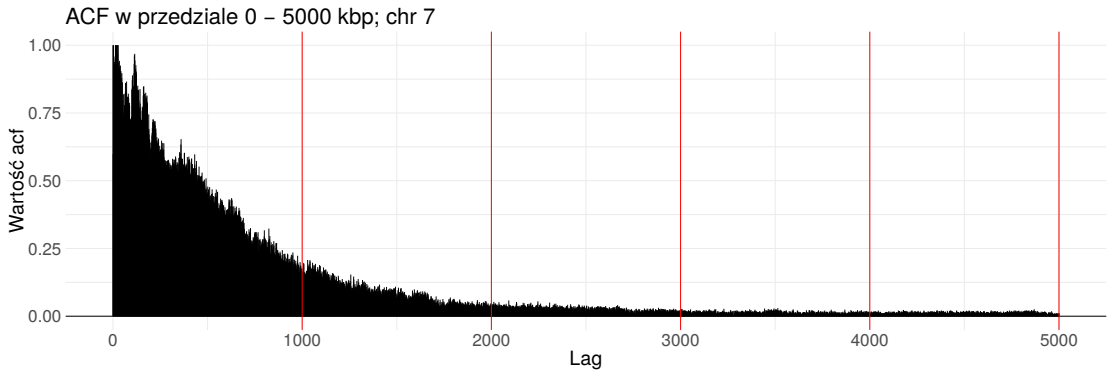
Rysunek 2.1: Stopień metylacji w dwóch próbach w chromosomie 2. Na osi X przedstawiono pozycję chromosomu, natomiast na osi Y stopień metylacji odpowiadającej danej pozycji. Kolor punktów wskazuje na logarytm głębokości sekwencjonowania. Pionową linią zaznaczono obserwacje z dwóch prób odnotowane na tej samej pozycji.

Rysunek 2.2 obrazuje zaś obszar, który nie jest interesujący.



Rysunek 2.2: Stopień metylacji w dwóch próbach w chromosomie 4.

Przedstawione wykresy 2.1 oraz 2.2 dobrze obrazują charakter danych i problemu. Oba z nich ilustrują obszary o długości 1000bp. Na obydwu można odnotować, że tylko na niektórych pozycjach DNA możliwy jest odczyt procesu metylacji lub jej braku. Widać zróżnicowanie sekwencjonowania, nawet dla obserwacji, które są położone blisko siebie. Pierwszy z zamieszczonych wykresów przedstawia obszar o dużym zróżnicowaniu metylacji. Można zauważyć, że intensywność metylacji w drugiej próbce ma tendencję do osiągania większych o około 0.4 wartości niż intensywność odnotowana w drugiej próbie. Kolejny z załączonych obrazków nie charakteryzuje się różnicą interesującą różnicą metylacji. Jest ona niewielka, w dodatku nie ma jednoznacznej przewagi intensywności metylacji na korzyść jednej z prób.
Przy opisie charakterystyki danych warto również zwrócić uwagę nie tylko na dużą korelację między próbkami ale również na bardzo duże wartości autokorelacji. Autokorelacja została obliczona na podstawie $\frac{meth+1}{unmeth+1}$, oddzielnie dla każdego chromosomu na podstawie odległości między pozycjami DNA. Rysunek 2.3 przedstawia wartości funkcji autokorelacji dla przykładowego chromosomu.



Rysunek 2.3: Wartość funkcji autokorelacji w chromosomie 7. Na osi X przedstawiono oddalenie szeregu intensywności metylacji a na osi Y odpowiadającą jej wartości korelacji. Pionowe linie odpowiadają kolejnym oddaleniom szeregu o 1000bp.

Duży poziom autokorelacji utrzymuje się nawet dla wartości oddalonych o 2000 bp. Dla wartości do 500 bp funkcja autokorelacji przekracza nawet 0.5.

2.2. Przegląd istniejących metod

2.2.1. Miary

Większość analiz dotyczących procesu metylacji odbywa się na podstawie ilości zmetylowanych i niezmetylowanych cytozyn oraz na dwóch zdefiniowanych uprzednio miarach. Jednym ze

wskaźników stosowanych w ilościowym określeniu procesu metylacji jest parametr β zdefiniowany poniżej (zob. Li 2015):

$$\beta = \frac{\max(M, 0)}{\max(M, 0) + \max(U, 0) + 100}, \tag{2.2}$$

gdzie:

M – liczba zmetylowanych cytozyn,

U – liczba niezmetylowanych cytozyn.

Miara służy do dalszego przetwarzania danych czyli m.in. do testów statystycznych czy wizualizacji . Wartość średnia parametru β będzie wskazywała na poziom metylacji w danym regionie. Naturalnie, wskazana miara przyjmuje wartości między 0 a 1, gdzie wartości bliskie 0 oznaczają brak lub niewielką ilość zmetylowanych cytozyn w stosunku do długości sekwencjonowania, natomiast wartości bliskie 1 będą wskazywały na sytuację odwrotną. Kolejna użyta w analizach miara to:

$$M = \log_2 \frac{\max(M, 0) + 1}{\max(U, 0) + 1}. \tag{2.3}$$

Przyjmuje ona wartości między $-\infty$ a ∞. Wartości bliskie 0 wskazują, że dany obszar został połowicznie zmetylowany. Wartości powyżej 0 mówią natomiast o większej ilości cytozyn zmetylowanych niż niezmetylowanych, zaś większa ilość cytozyn niezmetylowanych niż zmetylowanych będzie prowadziła do wartości poniżej 0. Zależność między obiema miarami może zostać opisana następująco:

$$M = log_2 \frac{\beta}{1 - \beta}. \tag{2.4}$$

Powyższe miary są szeroko stosowane w dotychczasowych analizach danych z badania metylacji. W tym celu stworzono kilka pakietów w środowisku statystycznym R. Są to m.in. *methyAnalysis* [1], *CpGassoc*[2], *RnBeads*[3], *IMA*[4] czy *minfi*[5]. Dostępne metody w wymienionych powyżej pakietach służące do znajdowania obszarów różnorodnie zmetylowanych to przede wszystkim: testy: test-t, Kołmogorowa-Smirnova, Wilcoxona, regresja liniowa, metoda empiryczna Bayesa oraz metoda bump-hunting. W większości wymienionych pakietów istnieje możliwość dostosowania wartości krytycznej na podstawie FDR (false discovery rate) dla ustalonego poziomu α.

[1] Du P. et al. *methyAnalysis: DNA methylation data analysis and visualization*, 2017.

[2] Barfield R. et al. *CpGassoc: Association Between Methylation and a Phentype of Interest*, 2017.

[3] Assenov Y. et al. *Comprehensive Analysis od DNA Methylation Data with RnBeads*, Nature Methods, 11(11):1138-1140, 2014.

[4] Wang D. et al. *IMA: an R package for high-through analysis of Ilumina's 450K Infiunium methylation data*, Bioinfmatics 28(5) s.729-30, 2017.

[5] Aryee M.J. et al. *Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays*, Bioinformatics 30(10), s.1363-1369, 2014.

2.2.2. Normalizacja danych

Dokładny opis metod służących analizie danych związanych z dziedziną metylacji należy rozpocząć od metod zajmujących się normalizacją. Odbywa się ona głównie poprzez normalizację kwantylową lub normalizację SWAN (Subset-quantile Within Array Normalization). Główną zaletą wspomnianych normalizacji jest otrzymanie danych, które mają takie same statystyczne własności. Przebieg normalizacji kwantylowej dla danych w postaci macierzowej, gdzie kolumny oznaczają wyniki metylacji kolejnych pacjentów a wiersze lokalizację chromosomu polega na:

1. Nadaniu rang poszczególnym wartościom osobno w każdej z kolumn
2. Uszeregowaniu rosnąco wartości w kolumnach i obliczenie średnich wierszowych dla danych poddanych transformacji. Otrzymane rezultaty będą utożsamiane z kolejnymi wartościami rang
3. Podstawienie wyników z pkt. 2 do rang zapamiętanych w pkt. 1

Dane z badania metylacji są często przetwarzane za pomocą narzędzia Infinium HumanMethylation450 BeadChip. Informacje pochodzące z działania tego narzędzia składają się z kombinacji próbek dwóch typów: Infinium I i Infinium II. Normalizacja danych metodą SWAN jest dobrym rozwiązaniem do normalizacji powyższego charakteru danych i bierze pod uwagę rodzaj próbki. Jest to istotne, ponieważ intensywności metylacji lub jej braku w obu typach próbek mają odmienny rozkład (zob. Maksimovic 2012). Procedura SWAN składa się z dwóch etapów. W pierwszym etapie obliczony zostaje średni rozkład kwantylowy na podstawie podzbioru biologicznie podobnych do siebie próbek. Zazwyczaj, próbki podobne biologicznie do siebie określa się poprzez równą ilość wysp CpG w każdej z nich. Dla tak wybranych próbek zostaje wykonana zwykła normalizacja kwantylowa w podziale na dane mówiące o intensywności procesu metylacji i jej braku. W następnym kroku dostosowane są intensywności pozostałych próbek poprzez liniową interpolację.

2.2.3. Uwzględnienie korelacji

Jednym ze sposobów radzenia sobie z wysoką korelacją obserwacji, które są blisko odległe od siebie oraz możliwym szumem technicznym jest uśrednienie poziomu metylacji na podstawie okna ruchomego o ustalonej długości. Często taki zabieg prowadzi do uzyskania takiego samego wyniku w sąsiednich pozycjach. Wynika to ze struktury danych. Odczyty poziomu metylacji są możliwe na określonych pozycjach. Przykładowo, odczyty na pozycji x oraz $x+k$ będą identyczne po zastosowaniu okna ruchomego o długości z, gdy między $[x-z, x+k–z)$ oraz $(x+z, x+z+k]$ nie zostaną zgromadzone dane.

2.2.4. Znajdowanie obszarów różnorodnie zmetylowanych

Pakiety statystyczne służące do analizy danych z badania metylacji przeprowadzają testy statyczne w oparciu o pozycje na każdym chromosomie oddzielnie. Odbywa się to poprzez porównanie wartości metylacji (β lub M zdefiniowane w 2.2 i 2.3) w dwóch grupach. Zazwyczaj jest to grupa objęta wskazaną chorobą oraz grupa kontrolna. Ponieważ dane na pobliskich lokalizacjach będą zazwyczaj identyczne po zastosowaniu okna ruchomego , uzyskane wartości krytyczne również będą tożsame. Definicja regionów różnorodnie zmetylowanych odbywa się najpierw poprzez określenie, czy wynik w danej pozycji jest istotny. Polega to głównie na sprawdzeniu czy wartość krytyczna lub wartości krytyczna po poprawce ze względu na FDR są mniejsze od ustalonych poziomów. Innym warunkiem, który również jest brany pod uwagę to różnica średniej lub mediany w obu grupach, przekraczająca określony próg. Następnie zostają zwrócone regiony maksymalnie odległe o określoną ilość pozycji, dla których wszystkie wyniki zostały oznaczone jako istotne. Opisany powyżej mechanizm odnosi się do podejścia wykorzystującego test-t na różnicę średnich (zakładający równą bądź nie wariancję), test na różnicę median Wilcoxona.
Kolejną z możliwych metod to regresja liniowa i wyznaczenie wartości krytycznej testu sprawdzającego istotność zmiennej wskazującej na podział między grupy. W tej metodzie zazwyczaj zakłada się, że dla g-tego geny mamy n wyników w postaci wektora odpowiedzi: $y_g^T = (y_{g1}, \cdots y_{gn})$, które spełniają (zob. Smyth 2004 s. 4-8):

$$E(y_g) = X\alpha_g, \tag{2.5}$$

$$var(y_g) = W_g\sigma_g^2, \tag{2.6}$$

gdzie:
W_g to macierz wag,
α_g - wektor współczynników,
X to macierz eksperymentu.

Niech C oznacza macierz kontrastów, wtedy współczynniki kontrastów można zapisać jako $\beta_g = C^T\alpha_g$. Kolejna wytyczna to:

$$var(\hat{\alpha}_g) = V_g\sigma_g^2, \tag{2.7}$$

gdzie:
V_g jest macierzą dodatnio określoną niezależną od σ_g^2.

Niech v_{gj} oznacza j-ty element na diagonali macierzy C^TV_gC. Założenia o rozkładach są wymienione poniżej:

$$\hat{\beta}_{gj}|\beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj}\sigma_g^2), \tag{2.8}$$

$$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2, \tag{2.9}$$

gdzie:

d_g to liczba stopni swobody w modelu liniowym dla genu g.

Zgodnie z powyższymi założeniami:

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g\sqrt{v_{gj}}} \tag{2.10}$$

ma w przybliżeniu rozkład t-Studenta z d_g stopniami swobody.

Modyfikacja powyższej statystyki bazuje empirycznym podejściu Bayesowskim, czyli na pewnych założeniach a priori co do rozkładu nieznanych parametrów. Zakłada się, że informacja zebrana a priori o σ_g^2 jest tożsama z estymatorem a priori s_0 z liczbą stopni swobody d_g, czyli (zob. Smyth 2004, s.8-9):

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi_{d_0}^2. \tag{2.11}$$

Powyższa zależność obrazuje zmianę wariancji w zależności od genu. Przyjmuje się również, że dla ustalonego j wartości współczynników β_{gj} są niezerowe z prawdopodobieństwem:

$$P(\beta_{gj} \neq 0) = p_j. \tag{2.12}$$

Wtedy p_j jest oczekiwaną proporcją różnorodnie zmetylowanych genów. Dla niezerowych współczynników zakłada się, że informacja zebrana a priori jest tożsama z obserwacją a priori równą 0 z nieprzeskalowaną wariancją v_{0j}, zatem:

$$\beta_{gj}|\sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2). \tag{2.13}$$

Powyższe równania opisują standardowy rozkład sprzężony do rozkładu normalnego opisanego uprzednio. Zgodnie modelem hierarchicznym wartość oczekiwana a posteriori parametru σ_g^2 na podstawie s_g^2 to:

$$\hat{s}_g^2 = E(\sigma_g^2|s_g^2) = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}. \tag{2.14}$$

Podstawiając nową wartość wariancji do statystyki z (2.10) otrzymuje się zmodyfikowaną statystykę testu t:

$$\hat{t}_{gj} = \frac{\hat{\beta}_{gj}}{\hat{s}_g\sqrt{v_{gj}}} \tag{2.15}$$

z $d_0 + d_g$ stopniami swobody.

Statystyka 2.15 reprezentuje połączenie klasycznego i Bayesowskiego podejścia. Zmodyfikowana statystyka t odzwierciedli standardowe podejście, gdy $d_0 = 0$.

Innym podejściem służącym do znajdowania obszarów o różnorodnym poziomie metylacji jest

test F zastosowany dla zmiennych kategorycznych w pakiecie *minfi*. Bazuje on na sumie wartości reszt z modelu liniowego oraz sumie wartości reszt w modelu liniowym wyłącznie ze stałą.

Na uwagę zasługuje również metoda *Bump hunting method* (zob. Jaffe 2012). Przebiega ona w kilku etapach. Na początku zostaje zdefiniowany model, np.:

$$Y_{ij} = \mu(t_j) + \beta(t_j)X_i + \sum_{k=1}^{p} \gamma_k(t_j)Z_{i,k} + \sum_{l=1}^{q} a_{l,j}W_{i,l} + \epsilon_{i,j}, \tag{2.16}$$

gdzie:

Y_{ij} oznacza stopień metylacji na j-tym lokusie dla i-tego pacjenta,

t_j - lokalizacja na genomie j-tego lokusa,

$\mu(t_j)$ -wartość średnia stopnia metylacji dla populacji,

X_i - macierz zmiennych objaśniających (np. podział na grupę chorą i kontrolną.),

$Z_{i,k}$ - potencjalne zakłócenia modelu (np. wiek, płeć nie będące głównym przedmiotem analizy),

$W_{i,l}$ - potencjalne, niezaobserwowane zakłócenia modelu,

$\epsilon_{i,j}$ - niewyjaśniona zmienność w modelu z $var(\sigma^2(tj))$ zależną od lokalizacji t_j.

Po zdefiniowaniu modelu następuje jego estymacja dla każdego j. Odbywa się to głównie za pomocą iteracyjnej metody SVA (zob. Jaffe 2012), gdyż zwykła regresja liniowa może nie być odpowiednia w przypadku braku informacji o macierzy W. Po uzyskaniu oszacowań $\hat{\beta}(t_j)$ następuje wysmuklenie funkcji $\hat{\beta(t)}$ uwzględniające błąd standardowy otrzymany z poprzedniego kroku. Interesującymi regionami będą regiony, dla których, po wysmukleniu $\beta(t) \neq 0 \quad \forall t \in R_n$. Ostatni etap to skorzystanie z testów permutacyjnych by osiągnąć statystyczną niepewność dla każdego estymowanego regionu.

2.3. Omówienie metod zastosowanych w pracy

2.3.1. Testy DMR

W pracy również posłużono się stopniem metylacji zdefiniowanym poprzez:

$$meth.rate_{tkj} = \frac{meth_{tkj}}{meth_{tkj} + unmeth_{tkj}}, \tag{2.17}$$

gdzie:

$meth_{tkj}$ to liczba zmetylowanych cytozyn,

$unmeth_{tkj}$ liczba niezmetylowanych cytozyn w chromosomie t na pozycji k i j-tej próbie.

W celu analizy, czy dany obszar ma różnicę metylacji istotną statystycznie najpierw następuje wybranie obserwacji, które na tej samej pozycji chromosomu posiadają wyniki w obu próbkach.

Następnie obliczony zostaje stopień metylacji, która wykorzystywany jest w testach t-Studenta, Wilcoxona i Kołmogorowa-Smirnowa. Konstruowane zostają dwa wektory:
$X = (meth.rate_{t_1k_11}; \ldots; meth.rate_{t_nk_n1})$, $Y = (meth.rate_{t_1k_12}; \ldots; meth.rate_{t_nk_n2})$
t.że $(t_i, k_i) \in A$, gdzie:
A oznacza zbiór par indeksów chromosomu i pozycji obszaru, dla którego przeprowadzany zostaje test statystyczny.

Przy użyciu testu t-Studenta wykorzystuje się statystykę testową zdefiniowaną w (1.3) dla par zależnych i sprawdza hipotezę dotyczącą wartości średniej różnicy intensywności metylacji:

$$H_0 : \mu_0 = 0, \quad H_1 : \mu_0 \neq 0. \tag{2.18}$$

Korzystając z testu Wilcoxona badana natomiast jest mediana różnicy intensywności metylacji:

$$H_0 : d = 0, \quad H_1 : d \neq 0. \tag{2.19}$$

Test Kołmogorowa-Smirnowa porównuje zaś rozkłady intensywności w dwóch próbach:

$$H_0 : F(x) = G(x) \text{ dla każdego x }, \quad H_1 : F(x) \neq G(x) \text{ przynajmniej dla jednego x.} \tag{2.20}$$

W przypadku modeli regresji logistycznych również w pierwszym kroku następuje wybranie obserwacji, które na tej samej pozycji chromosomu posiadają wyniki w obu próbach. Nie jest jednak koniecznie obliczanie stopnia metylacji, ponieważ wykorzystywana jest bezpośrednio informacja o ilości zmetylowanych bądź nie cytozynach. We wszystkich trzech przypadkach aplikowany zostaje model danych zgrupowanych, gdzie metylacja cytozyny oznacza sukces natomiast długość sekwencjonowania - liczbę wszystkich zdarzeń. Modele regresji logistycznej mają jedną zasadniczą zaletę w porównaniu do testów statystycznych bazujących na intensywności. Tak jak wspomniano, bazują one na liczbie zmetylowanych i niezmetylowanych cytozyn. Dzięki temu nie jest tracona informacja o dokładnym charakterze danych, która ma miejsce przy analizowaniu jedynie intensywności.

O rozkładzie $meth_{tkj}$ zakłada się:

$$meth_{tkj} \sim Bin(meth_{tkj} + unmeth_{tkj}, \pi_i). \tag{2.21}$$

Dokładny model dla standardowej regresji logistycznej wygląda następująco:

$$logit(\pi_i) = \beta_0 + \beta_1 \cdot g_{1i} + \sum_{k \in K, s \in 2,\ldots,n} \beta_s \cdot p_{ik}, \tag{2.22}$$

gdzie:
g_{1i} oznacza indykator, czy i-te obserwacje pochodzą z 1. próbki,
p_k to indykatory, czy i-ta obserwacja dotyczy k-tej pozycji w obszarze który jest testowany, gdzie

pozycja o najmniejszym rankingu stanowi poziom odniesienia,
n to liczba pozycji dla których zgromadzono dane w testowanym obszarze.

Po przeprowadzeniu estymacji parametrów następuje sprawdzenie czy obie próbki różnią się statystycznie istotnie poziomem metylacji. Przeprowadzany zostaje test dla parametru β_1:

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0, \tag{2.23}$$

gdzie hipoteza zerowa zostaje odrzucona na podstawie testu Walda. Najpierw wylicza się:

$$z = \frac{\hat{\beta}_1}{se(\beta_1)} \tag{2.24}$$

a następnie wartość statystyki zostaje przybliżona rozkładem normalnym. Ostatecznie wartość krytyczna wynosi:

$$2 \times \phi(\frac{-|\hat{\beta}_1|}{SE(\beta_1)}). \tag{2.25}$$

Należy zauważyć, że w standardowym modelu regresji logistycznej zasadnicza większość parametrów: β_0; $\beta_2; \ldots; \beta_n$ jest estymowana ale nie poddawana dalszej analizie. Wymienione parametry nie wnoszą informacji czy obie próbki różnią się poziomem metylacji. Powstaje pytanie, czy istnieje potrzeba ich estymacji. Z pomocą przychodzą modele mieszane. Korzystając z modeli mieszanych otrzymuje się mniejszą ilość parametrów do oszacowania zatem dopasowanie modelu powinno być bardziej dokładne. W przypadku modeli mieszanych rozkład warunkowy $meth_{tkj}$ jest analogiczny jak w standardowej regresji logistycznej:

$$meth_{tkj}|u \sim Bin(meth_{tkj} + unmeth_{tkj}, \pi_i). \tag{2.26}$$

Natomiast model mieszany może zostać zapisać poprzez:

$$logit(\pi_i) = \beta_0 + \beta_1 \cdot g_{1i} + p_i u, \tag{2.27}$$

gdzie:
u oznacza wektor efektów losowych,
p_i to wektor indykatorów $p_{ik}, k \in K$ dla i-tej obserwacji.

W pracy rozważano dwie struktury macierz kowariancji efektów losowych. Pierwsza wersja dotyczy przypadku, gdy $u \sim N(0, \sigma^2 D(\theta))$ gdzie $D(\theta) = I$. Powyższy przypadek zakłada więc brak korelacji między efektami losowymi, czyli kolejnymi pozycjami na chromosomach. Estymacja współczynników modelu odbyła się poprzez maksymalizację marginalnej funkcji wiarygodności przybliżonej metodą aproksymacji Laplace'a.

Drugi przypadek uwzględniał strukturę autokorelacji na poszczególnych pozycjach. Skorzystano z macierzy kowariancji, która opierała się o średnią funkcję autokorelacji między poszczególnymi chromosomami. Tabela 2.3 zawiera przykład macierzy korelacji efektów losowych.

pozycja	10	11	20	21
10	1.00	0.72	0.84	0.88
11	0.72	1.00	0.88	0.84
20	0.84	0.88	1.00	0.72
21	0.88	0.84	0.72	1.00

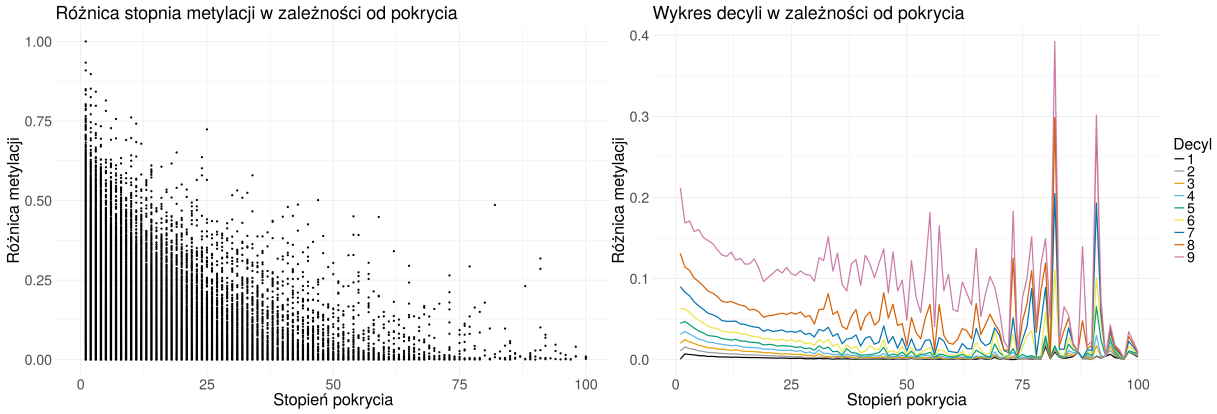
Tablica 2.3: Przykładowa macierz korelacji

Warto zauważyć, że wartość korelacji między pozycją 10. i 11. oraz 20. i 21. ma taką samą wartość. Wynika to z tego, że te obserwacje są od siebie równoodległe.

W tym przypadku estymacja opierała się o maksymalizację PQL.

2.3.2. Regresja kwantylowa

W celu porównania wyników poszczególnych metod posłużono się regresją kwantylową. Dzięki niej przyporządkowano odpowiedni kwantyl dla różnicy stopnia metylacji w danym regionie. Przyporządkowany kwantyl będzie pełnił rolę niejakiego rankingu: im bliższe wartości 1, tym wybrany obszar będzie bardziej obiecujący w dalszej analizie.
Rysunek 2.4 przedstawia stopień metylacji oraz decyle próbkowe w zależności od pokrycia.



(a) Rozkład różnicy metylacji.

(b) Rozkład decyli różnicy metylacji.

Rysunek 2.4: Charakterystyki rozkładu różnicy metylacji. Po lewej rozkład różnicy metylacji: na osi X przedstawiono stopień pokrycia obszaru, na osi Y różnicę średniego poziomu metylacji w dwóch grupach. Po prawej rozkład decyli różnicy metylacji: na osi X również przedstawiono stopień pokrycia obszaru, na osi Y zaś kolejne decyle średniego poziomu metylacji estymowane oddzielnie dla każdego stopnia pokrycia obszaru.

Na powyższym wykresie można zauważyć, że przeważają obszary o pokryciu mniejszym niż 30 obserwacji. Im mniej obserwacji w danym regionie tym częściej obserwowane są większe różnice metylacji. Sytuacja zmienia się dla obszarów o pokryciu powyżej 30 jednostek. Liczba regionów

zmniejsza się co prowadzi do dużej niestabilności oszacowań kolejnych decyli.

W celu pozbycia się niestabilności oszacowań w regresji kwantylowej zdecydowano się na estymację współczynników wyłącznie dla regionów zawierających do 30 obserwacji. Dla obszarów o większej liczbie obserwacji skorzystano z oszacowanych parametrów. Szczegółowa estymacja wyglądała następująco:

1. Estymacja dla $l = 0.01, \ldots, 0.99$:

$$\hat{\beta}_l(\tau = l) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum \rho_\tau (y_i - \beta_0 - f(x_i)'\beta_1)^2,$$

   gdzie:

   y_i oznacza różnicę metylacji w i-tym obszarze,

   x_i -pokrycie i-tego obszaru,

   f to funkcja przekształcająca pokrycie obszaru.

   Zdecydowano się na $f(n) = \frac{1}{\sqrt[3]{n}}$, gdyż najlepiej odzwierciedlała postać danych i poprawiała dopasowanie modelu w porównaniu do $f(n) = \frac{1}{n}$ czy $f(n) = n$.

2. Następnie dla wartości $p = 1, \ldots, 300$:

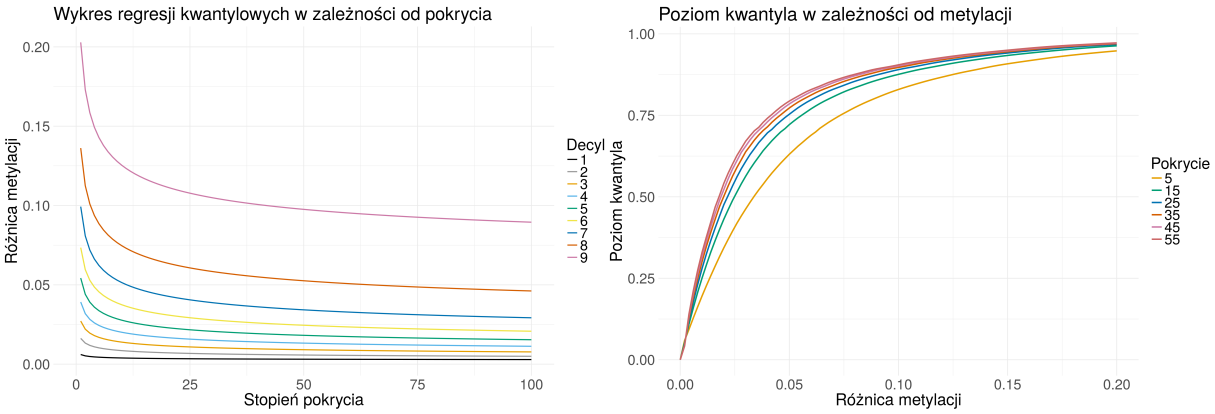
   Dla każdego $l = 0,01; \ldots; 0,99$ zostaje obliczony:

$$q_{pl} = \beta_{0l} + f(p)\beta_{1l},$$

   który jest oszacowaniem kwantyla o poziomie l dla regionu o pokryciu równym p.

3. Dla każdego p, mając ciąg wartości $q_{p;0,01}, \ldots, q_{p;0,99}$ (czyli kolejnych kwantyli) można dokonać liniowej aproksymacji pamiętając o dwóch ważnych ograniczeniach: kwantyl na poziomie 0 wynosi 0, natomiast na poziomie 1 to 1.

W wyniku wyżej opisanej procedury otrzymano funkcję, która niezależnie dla każdego obszaru o zadanym stopniu pokrycia zwróci wartość kwantyla odpowiadającego różnicy stopnia metylacji. Składa się ona z p niezależnych aproksymacji liniowych dla każdego regionu o danym pokryciu z przedziału $\{1; \ldots; p\}$.



(a) Regresja kwantylowa w zależności od pokrycia. (b) Poziom f. rankingowej w zależności od pokrycia.

Rysunek 2.5: Wynik zastosowania regresji kwantylowej. Po lewej wykres regresji kwantylowej w zależności od pokrycia: na osi X przedstawiono stopień pokrycia obszaru, na osi Y kolejne decyle różnicy średniego poziomu metylacji estymowane na podstawie regresji kwantylowej. Po prawej wykres funkcji kwantylowej: na osi X przedstawiono różnicę średniego poziomu metylacji, na osi Y poziomy funkcji rankingowej aproksymowanej osobno dla każdego stopnia pokrycia obszaru.

Z przedstawionych wykresów można odczytać, że największa różnorodność decyli występuje dla obszarów o bardzo małym pokryciu. Warto również zwrócić uwagę na zależność między funkcją rankingową a stopniem pokrycia. Dla danej różnicy średniego poziomu metylacji, im większy stopień pokrycia, tym funkcja rankingowa osiąga większe wartości.

3. Porównanie metod

3.1. Opis symulacji

Metoda 1: Metoda, która ma za zadanie odtworzyć naturalny rozkład danych. Przebiega identycznie i niezależnie dla każdej z badanych grup:

1. Z każdej pozycji k chromosomu t i próbki j pobrana zostaje informacja o stopniu metylacji $meth.rate_{tkj}$
2. Następnie, dla każdej pozycji została wylosowana liczba cytozyn, które uległy metylacji $meth_{tkj}$ zgodnie z rozkładem Bernoulliego $\sim Bern(\overline{meth_{tkj}} + \overline{unmeth_{tkj}}, meth.rate_{tkj})$. Prawdopodobieństwo sukcesu w tym rozkładzie było równe rzeczywistemu stopniu metylacji a liczba zdarzeń stanowiła sumę orginalnie zmetylowanych ($\overline{meth_{tkj}}$) i niezmetylowanych cytozyn ($\overline{unmeth_{tkj}}$).
3. Zgodnie z powyższym liczba sukcesów będzie wskazywała na liczbę zmetylowanych cytozyn a różnica między liczbą zdarzeń a liczbą sukcesów na liczbę cytozyn, które nie uległy metylacji.

Metoda 2: Powyższa metoda zakłada, że 5% analizowanych regionów jest różnorodnie zmetylowana, natomiast pozostałe 95% regionów nie wykazuje różnic w stopniu metylacji. Symulacje odbywają się jednocześnie dla obu grup w następujący sposób:

1. Próbka została podzielona na regiony o długości 1000bp. Z każdego regionu został pobrany średni stopień metylacji w obu grupach ($meth.rate_r$)
2. Kolejno, losowo podzielono wszystkie regiony na grupy o liczności 95% (grupa 1.) i 5% (grupa 2.)
3. Losowanie liczby zmetylowanych cytozyn - $meth_{tkj}$ również odbywało się zgodnie z rozkładem Bernoulliego. W grupie 1., dla obu próbek: $Bern(\overline{meth_{tkj}} + \overline{unmeth_{tkj}}, meth.rate_r)$. Natomiast w grupie 2., pierwsza próbka została wylosowana zgodnie z rozkładem: $Bern(\overline{meth_{tkj}} + \overline{unmeth_{tkj}}, meth.rate_r)$ natomiast druga z: $Bern(\overline{meth_{tkj}} +$

$\overline{unmeth_{tkj}}, meth.rate_r + p)$, gdzie p to dodatkowy parametr mówiący o różnicy metylacji w obu próbkach.

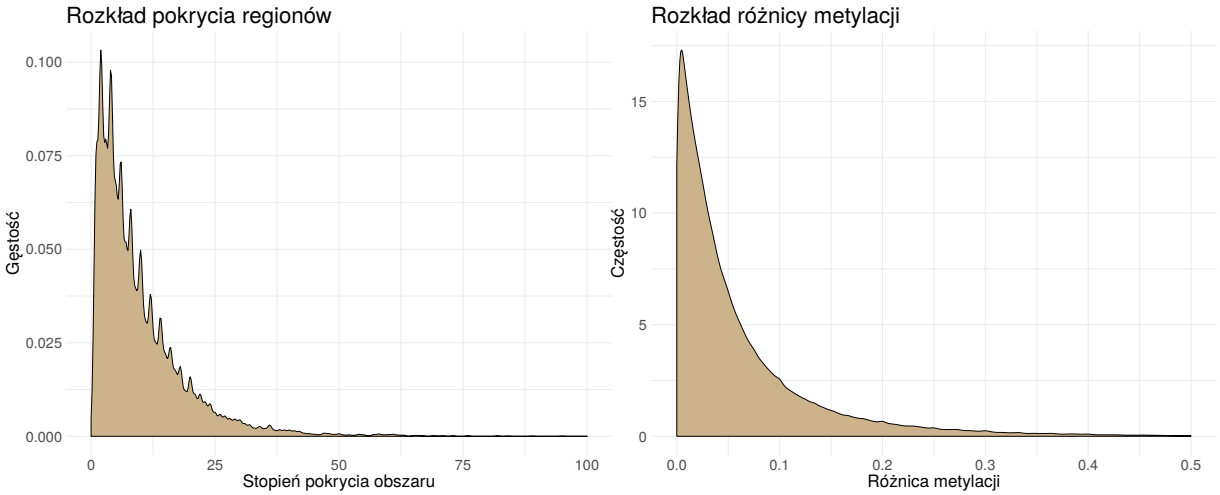
4. Również w tej metodzie liczba sukcesów w rozkładzie Bernoulliego będzie wskazywała na liczbę zmetylowanych cytozyn a różnica między liczbą zdarzeń a liczbą sukcesów na liczbę cytozyn, które nie uległy metylacji

W obu metodach zachowano rzeczywistą wartość pokrycia. Zmianie uległ jedynie stopień metylacji.

3.2. Wyniki symulacji

3.2.1. Metoda 1.

Rysunek 3.1 prezentuje rozkład pokrycia i różnicy metylacji w obu próbach. Wykresy są zbliżone do rzeczywistych danych, na których bazują parametry do przeprowadzenia symulacji.



(a) Rozkład pokrycia

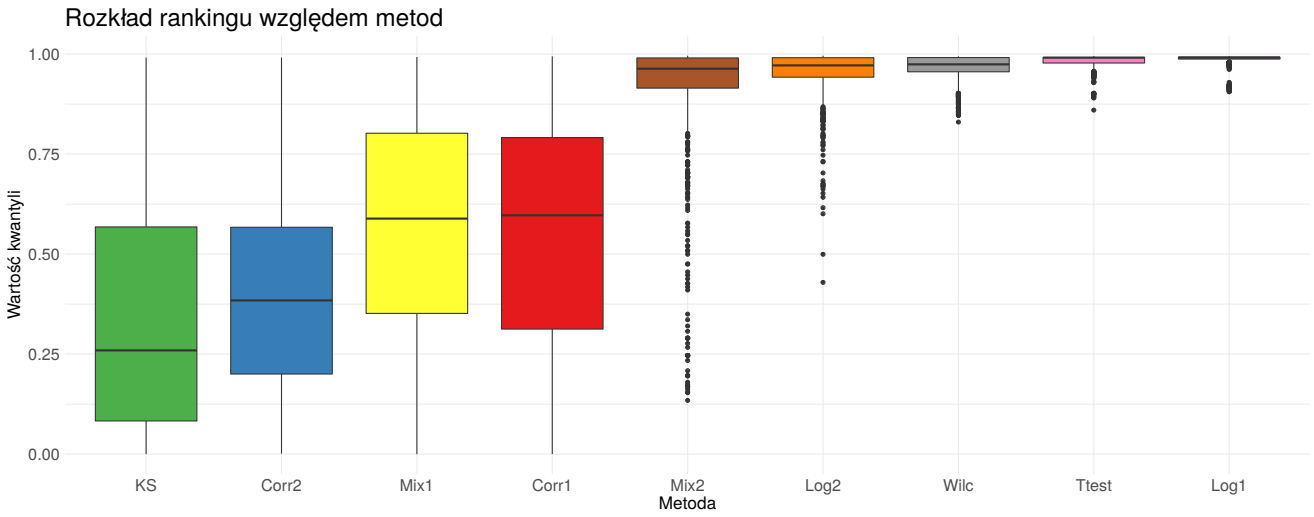
(b) Rozkład różnicy metylacji

Rysunek 3.1: Rozkłady danych pochodzących z symulacji w metodzie 1. Po lewej rozkład pokrycia regionów: Przedstawiono gęstość stopnia pokrycia regionu, czyli liczbę obserwacji objętej tym samym testem statystycznym w jednej próbie. Po prawej rozkład różnicy metylacji: przedstawiono gęstość wartości absolutnej różnicy średnich stopni metylacji między próbami w danych regionie.

Dane pochodzące z symulacji zostały uprzednio podzielone na regiony o długości 1000bp za pomocą okna ruchomego. W jednym obszarze znajdują się obserwacje, których pozycja na chromosomie po podzieleniu całkowitym przez 1000 daje identyczny wynik. Pokrycie danego

obszaru mówi o ilości obserwacji w dwóch grupach, natomiast różnica metylacji to wartość bezwzględna z różnicy średnich w obu grupach. W sytuacji podziału na regiony o długości 1000bp przeważają obszary o pokryciu do 10 obserwacji. Spośród 653800 regionów skonstruowanych na podstawie pierwszej metody jedynie 0.67 % to regiony o pokryciu powyżej 50. W przypadku różnicy metylacji również widoczna jest prawoskośność rozkładu. Przeważają obszary o niewielkich rozbieżnościach w średnich metylacjach. Różnica powyżej 0.1 dotyczy 18.5 % obszarów a powyżej 0.15 już tylko 10.1 % obszarów.

Wykres 3.2 obrazuje rozkład rankingu względem zastosowanych podejść. Ranking bazował na wynikach regresji kwantylowej, która miała na celu estymację rozkładu różnicy średnich metylacji (omówiony dokładnie w rozdziale 2.3.2.). Przyjmuje on wartości z przedziału $[0, 1]$. Wartości rankingu bliskie 1 świadczą o bardzo dużej różnicy średnich, natomiast bliskie 0 o znikomej różnicy średnich spośród obszarów o danym pokryciu.

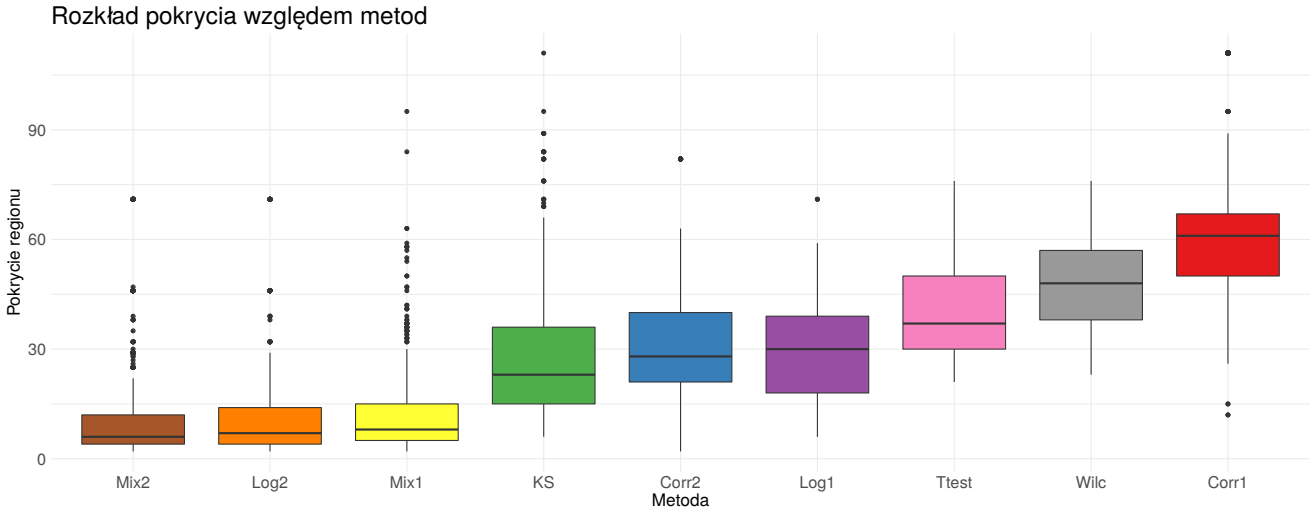


Rysunek 3.2: Rozkład rankingu top1000 obszarów wybranych względem zastosowanych podejść. Na osi X przedstawiono uszeregowane rosnąco względem mediany rankingu badane metody, na osi Y zaś rozkłady rankingów za pomocą box-plotów.

Umieszczony rysunek wskazuje na duże różnice względem metod, które zostały poddane analizie. Najgorzej radzi sobie podejście bazujące na statystyce Kołmogorowa-Smirnowa (KS). Mediana uzyskanego rankingu jest niewiele większa od 0.25, ponadto jest to podejście bardzo niestabilne, wskazujące zarówno obszary o niewielkiej jak i dużej różnicy metylacji. Sposób znajdowania DMRów na podstawie regresji logistycznej z efektami losowymi uwzględniającej strukturę autokorelacji również zawodzi. Obszary wybrane na podstawie uszeregowania rosnąco wartości krytycznych (Corr1) czy uszeregowania malejąco wartości zmiennej grupującej (Corr2) są również bardzo zróżnicowane. Rzadko prowadzą do wskazania regionów, dla których ranking przyjmuje wartości bliskie 1. Ciekawą obserwacją jest również fakt, że metoda szeregująca ob-

szary na podstawie wielkości efektu zmiennej grupującej prezentuje się gorzej niż metoda biorąca pod uwagę jedynie istotność zmiennej grupującej. Odwrotny wniosek wysuwa się gdy porównane zostaną podejścia bazujące na regresji logistycznej z efektami losowymi. W tym przypadku o wiele lepsze rezultaty są widoczne dla metody szeregującej obszary na podstawie wielkości efektu (Mix2) niż w oparciu o wartości krytyczne (Mix1). Bardzo duża część obszarów wybranych za pomocą Mix2 posiada ranking bliski 0.95, jednak warto również zwrócić uwagę na obserwacje odstające. Są wśród nich obszary, które nie są interesujące do analizy ponieważ rankingi dla nich są poniżej 0.75. Biorąc pod uwagę wartości klasyfikacji na podstawie regresji kwantylowej na wyróżnienie zasługują trzy metody: podejście bazujące na teście t-Studenta (Ttest), teście Wilcoxona (Wilc) oraz regresji logistycznej. Mediana rankingów uzyskanych przez te metody jest większa od 0.98. Najmniejszą zmienność ale również i najmniejszą liczbę wartości odstających posiadają regiony wybrane na podstawie regresji logistycznej szeregującej obszary na podstawie wartości krytycznych (Log1). Nieznacznie gorsze obszary zostały wskazane przez test t-Studenta i Wilcoxona. Podobny wynik został osiągnięty na podstawie regresji logistycznej wykorzystującej wielkość efektu (Log2), jednak ta metoda charakteryzuje się większą liczbą obszarów odstających w porównaniu do trzech metod, które wskazują najbardziej interesujące obszary.

Wykres 3.3 przedstawia rozkład pokrycia 1000 najbardziej interesujących obszarów wybranych poprzez 9 analizowanych metod.

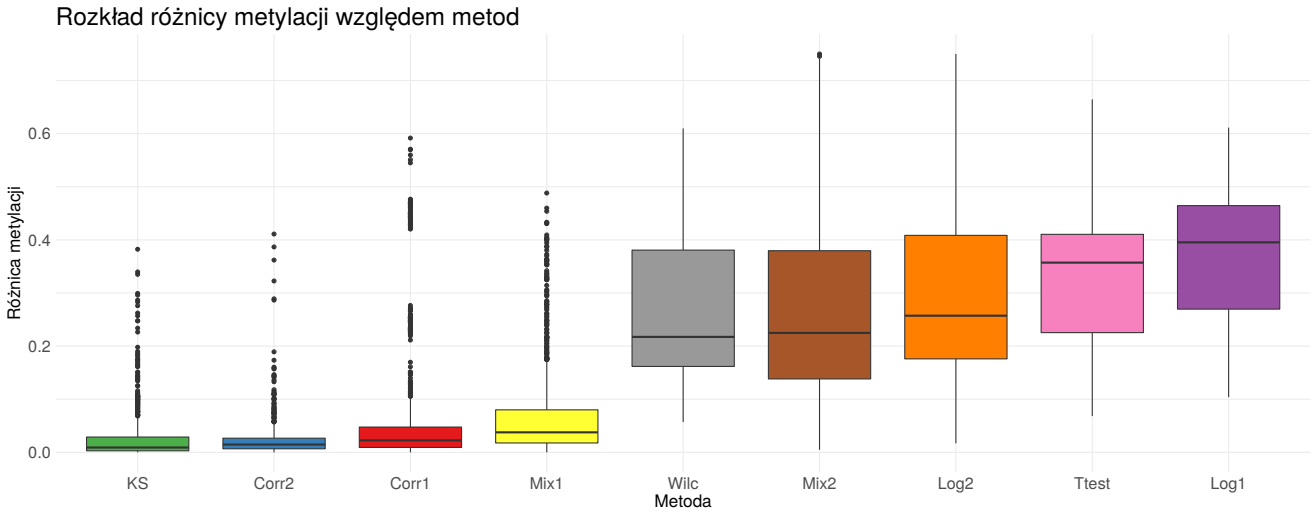


Rysunek 3.3: Rozkład pokrycia top1000 obszarów wybranych względem zastosowanych podejść. Na osi X przedstawiono uszeregowane rosnąco względem mediany pokrycia badane metody, na osi Y zaś rozkłady pokryć za pomocą box-plotów.

Przedstawiony rysunek wskazuje na różnicę względem działania poszczególnych metod. Każda z nich rekomenduje obszary charakteryzujące się odmiennym rozkładem pokrycia. Warto zauważyć, że metody bazujące na regresjach logistycznych (Corr, Mix, Log) posługujące się jedynie

wartościami krytycznymi wskazują zawsze na obszary o większym pokryciu niż metody bazujące na wielkości efektu zmiennej grupującej. Różnica jest widoczna przede wszystkim w przypadku metod Corr i Log. Spośród pięciu metod, które wskazywały na regiony o największym zróżnicowaniu stopnia metylacji w dwóch grupach największe pokrycie miała grupa regionów wybrana przez test Wilcoxona. Były to obszary zazwyczaj o pokryciu powyżej 40 obserwacji per region. Odrobinę mniejszym pokryciem, powyżej 30 charakteryzują się obszary rekomendowane przez test t-Studenta. Metodę Log1 można asocjować z interesującymi obszarami oscylującymi w granicach 16-34 obserwacji. Powyższe trzy metody mają większą różnorodność niż metody Log2 i Mix2. Te procedury wskazują na regiony o pokryciu poniżej 15. Mniejsza różnorodność aparatu Log2 i Mix2 może wynikać z faktu, że regiony poniżej 15 obserwacji przeważają w próbie. Metody cechujące się polecaniem regionów o niskim pokryciu mają więc wystarczająco dużo obszarów, by wśród nich wybrać te najbardziej interesujące.

Rysunek 3.4 przedstawia rozkład różnicy metylacji dla top100 między metodami.

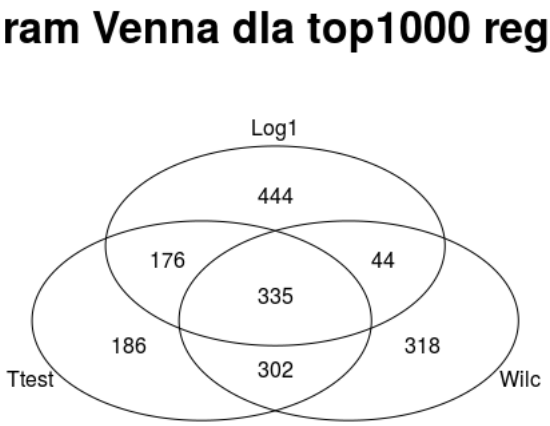


Rysunek 3.4: Rozkład różnicy metylacji top1000 obszarów wybranych względem zastosowanych podejść. Na osi X przedstawiono uszeregowane rosnąco względem mediany różnicy stopnia metylacji badane metody, na osi Y zaś rozkłady pokryć za pomocą box-plotów.

Pięć metod, które osiągnęły najlepiej rokujące wyniki rankingu posiada również największe różnice absolutne wartości średnich stopni metylacji w dwóch próbach. Można zauważyć większą dysproporcję między medianami różnicy metylacji niż dla median rankingu. Wynika to z faktu, że ranking uwzględnia oddzielnie rozkłady o różnych pokryciach. Największe różnice metylacji występują w grupie obszarów o niewielkim pokryciu. W związku z tym, że test Wilcoxona ma tendencję do rekomendacji obszarów o stosunkowo większej ilości obserwacji (osiąga tam mniejsze wartości krytyczne), wybiera on regiony o relatywnie mniejszej różnicy metylacji. Metody bazujące na wielkości efektu (Mix2, Log2) wybierają obszary zbliżone do metody Wilc pod

względem różnicy metylacji. Aparat Log1 wybiera natomiast obszary, gdzie różnica metylacji jest największa, co prowadzi do selekcji rejonów o niewielkiej ilości obserwacji. Test t-Studenta rekomenduje zaś dość liczne obszary o stosunkowo dużej różnicy metylacji.

W związku z tym, że każda z metod wskazuje na regiony o różnym stopniu pokrycia oraz odmiennej różnicy metylacji, istnieje bardzo mała liczba obszarów wybrana wspólnie przez kilka podejść. Obrazek 3.5 przedstawia Diagram Venna dla top1000 obszarów wybranych przez 3 najlepsze podejścia pod względem rankingu.



Rysunek 3.5: Diagram Venna dla trzech najlepszych metod wybranych na podstawie mediany rankingu kwantylowego.

Około 33 % regionów to regiony rekomendowane wspólnie przez trzy najlepsze metody. Największe podobieństwo w mechanizmie wybierania DMRów wykazują test Wilcoxona i t-Studenta. Łącznie, ponad 66 % regionów znajdujących się w top1000 zostało wybranych obiema metodami. Regresja logistyczna wykazuje największą autonomiczność, posiada ok. 38% regionów wspólnych z testem Wilcoxona i 51% obszarów wybranych przez test t-Studenta.

Porównując podejścia służące do znajdowania DMRów warto również zwrócić uwagę na ilość obszarów, które nie mogły zostać przetworzone. Wynika to z niewystarczającej ilości obserwacji umożliwiającej przeprowadzenie testu t-Studenta lub Wilcoxona czy poprzez niezbieganie algorytmu obliczającego estymatory największej wiarygodności w przypadku regresji logistycznych. Szczegółowe informacje zostały zamieszczone w tabeli 3.1.

Metoda	Ttest	Wilc	KS	Log1	Mix1	Corr1
Lp. nieprz. obszarów	54526	13537	0	48591	62061	116877
% nieprz. obszarów	8.34%	2.07%	0%	7.43%	9.49%	17.88%

Tablica 3.1: Porównanie liczby nieprzetworzonych obszarów w podziale na metody.

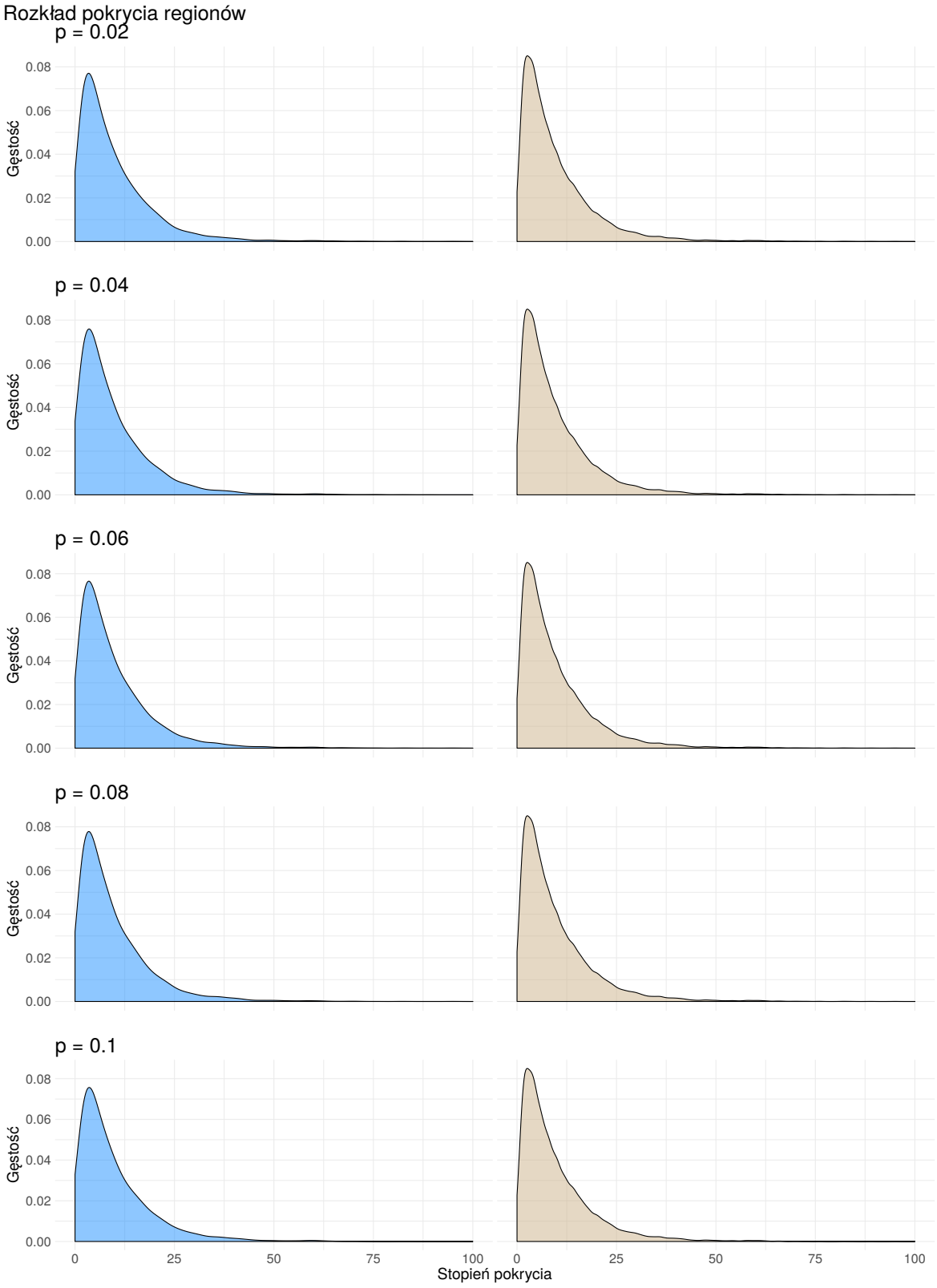
Tylko jedna metoda była w stanie przetworzyć wszystkie obszary, jest to test Kołmogorowa-

Smirnowa. Niewielka liczba nieprzetworzonych obszarów dotyczy również testu Wilcoxona. Największa liczba regionów, dla których nie uzyskano wyniku odnosi się do metody Corr i stanowi 17.88 % całości danych. Rekomendacje na podstawie wielkości efektu pochodzących z regresji logistycznych mają wyniki tożsame z ich odpowiednikami bazującymi na sortowaniu obszarów po wartościach krytycznych zmiennej grupującej. W dalszym kroku dla Log2, Mix2, Corr2 wybierane są obszary do dalszej analizy, dla których wartość krytyczna zmiennej grupującej jest mniejsza od 0.001. a następnie sortowane względem wartości oszacowania zmiennej grupującej.

3.2.2. Metoda 2.

Symulacje na podstawie drugiego podejścia bazują na dodatkowym parametrze p. Rysunek 3.6 przedstawia rozkład pokrycia regionów w zależności od przyjętego parametru p.

Rozkład pokrycia w symulacji metodą 2. jest podobny do rozkładu pokrycia w symulacji metodą 1., ponieważ bazuje na identycznych danych na temat pokrycia. Niezależnie od parametru p liczba regionów poddanych analizie wynosi 392 280. 19 620 (5 %) z nich zostało zasymulowane zgodnie z prawdopodobieństwem sukcesu równym średniej metylacji w danym obszarze + parametr p w wybranej grupie. Warto zauważyć, że odsetek regionów z zwiększonym parametrem symulacji o p w grupie regionów o podobnym pokryciu jest zbliżony. Dzięki temu metody nie będą a priori faworyzowały grupy regionów z wybranym pokryciem, w którym zasadnicza grupa regionów charakteryzuje się dużą różnicą stopnia metylacji.



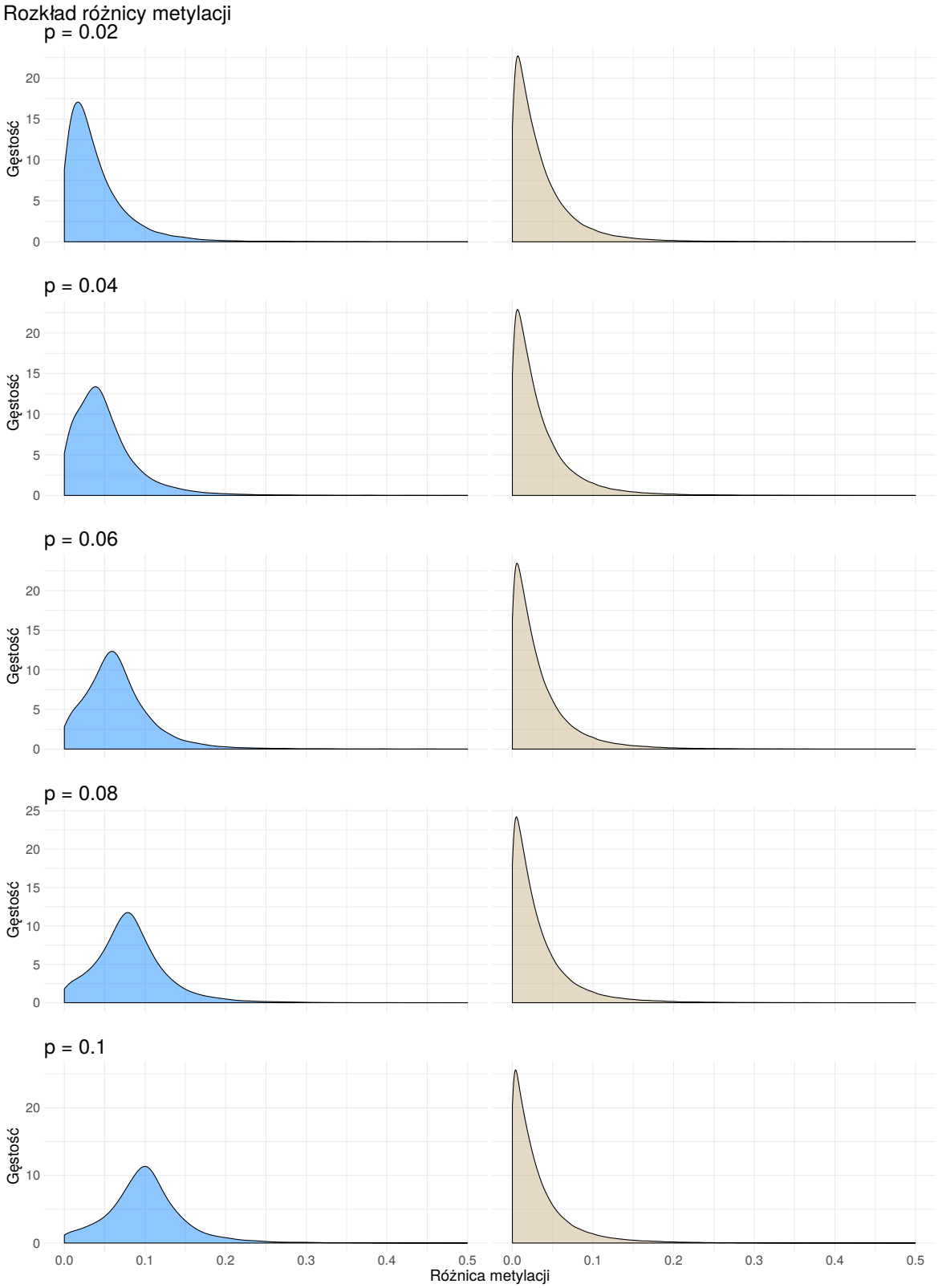
Rysunek 3.6: Rozkład pokrycia obszarów w zależności od parametru p. Wykresy po lewej przedstawiają rozkłady regionów z dodatkowym parametrem symulacji, zaś po prawej bez dodatkowego parametru.

Rysunek 3.7 przedstawia rozkład różnicy metylacji, zdefiniowanej analogicznie jak w przypadku symulacji w części 1. Warto zauważyć, że podobnie jak w symulacjach w części pierwszej większość obszarów to obszary o różnicy metylacji poniżej 0.1. Można również dostrzec, że zwiększając parametr p, liczba obszarów o różnicy powyżej 0.1 również wzrasta. Szczegółowe informacje można uzyskać w poniższej tabeli.

Metoda	p = 0.02	p = 0.04	p = 0.06	p = 0.08	p = 0.1
Lp. obszarów z różnicą met. > 0.1	25011 (1391)	25382 (1908)	26224 (3174)	27974 (5373)	31843 (9475)
% obszarów z różnicą met. > 0.1	6.38 % (5.56 %)	6.47 % (7.52 %)	6.69 % (12.1 %)	7.13 % (19.21 %)	8.11 % (29.75 %)

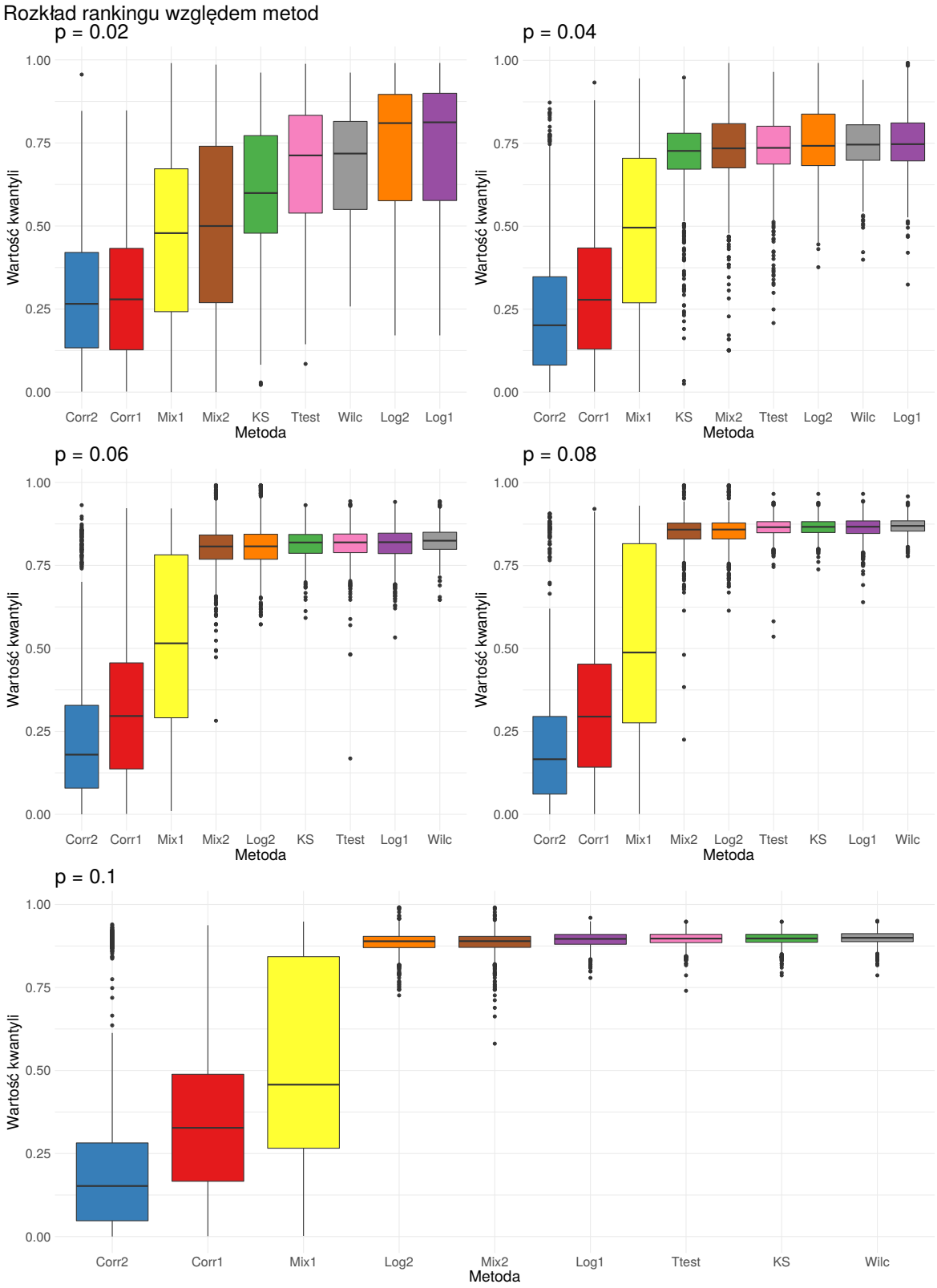
Tablica 3.2: Porównanie liczby obszarów w podziale na parametr symulacji p

W nawiasach podano wartości w przypadku regionów, dla których symulacje odbywały się z prawdopodobieństwem sukcesu równym średniej metylacji w danym obszarze + parametr p w wybranej grupie. Wartości procentowe w tym przypadku odnoszą się do do wszystkich regionów, które posiadały dodatkowy parametr symulacji. Rosnąca liczba regionów o różnicy metylacji powyżej 0.1 wraz ze wzrostem parametru p wynika głównie z rosnącej liczby regionów, gdzie parametry symulacji różniły się w obu grupach.

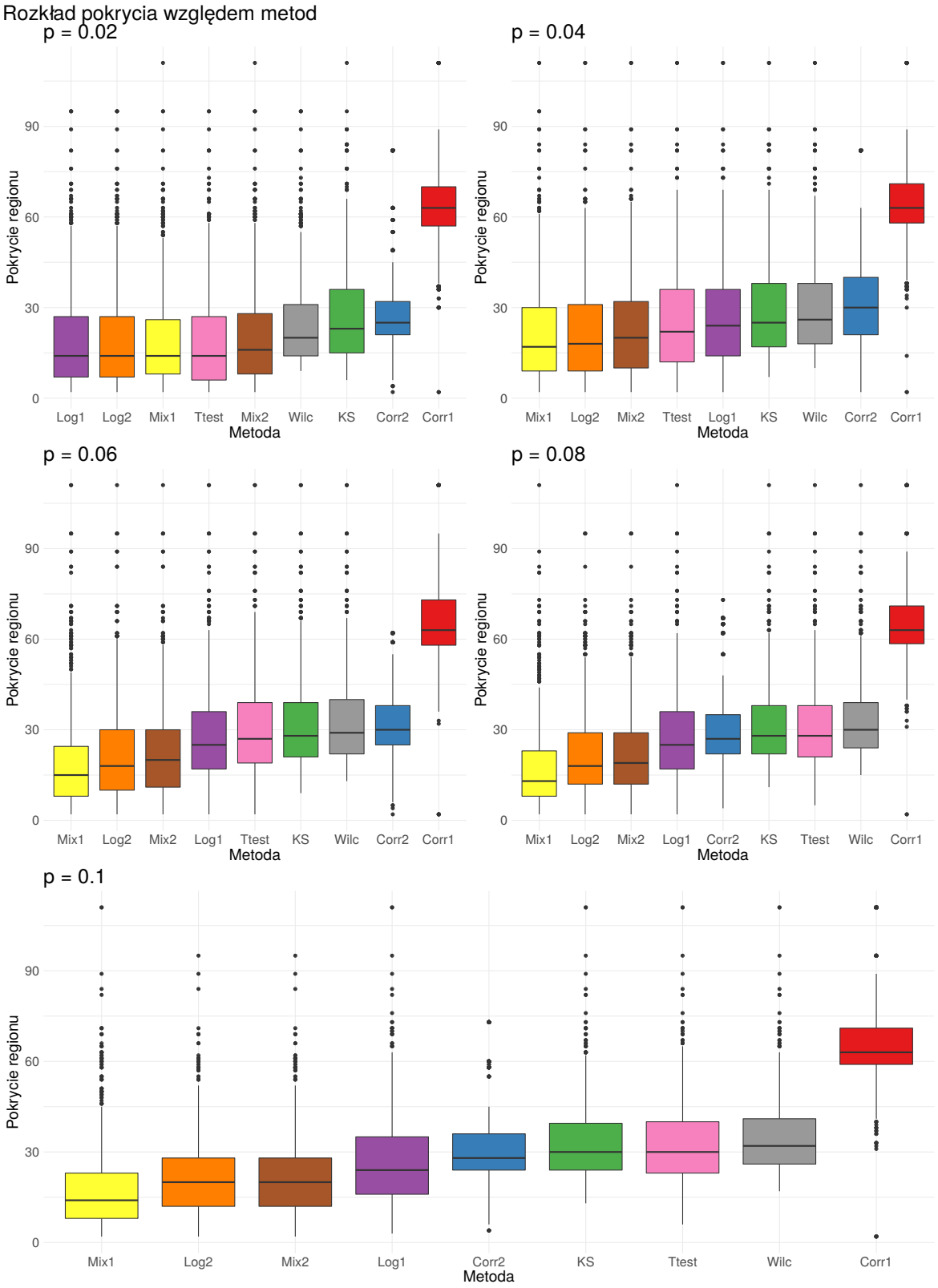


Rysunek 3.7: Rozkład różnicy metylacji w zależności od parametru p. Wykresy po lewej przedstawiają rozkłady różnicy metylacji z dodatkowym parametrem symulacji, zaś po prawej bez dodatkowego parametru.

Wykresy 3.8 prezentują rozkład rankingu top1000 obszarów wybranych przez 9 analizowanych metod. Najmniej korzystne wyniki zostały otrzymane poprzez regresję logistyczną z efektami losowymi uwzględniającą korelację danych. Widoczna jest duża zmienność obu metod Corr1 i Corr2 we wszystkich pięciu przypadkach. Dostrzec można niewielką poprawę na rzecz wzrostu mediany rankingów w sytuacji wzrostu parametru p, nie jest to jednak tak spektakularny wzrost jak w przypadku pozostałych metod. Procedura Corr2 wskazuje na obszary o niższym rankingu niż procedura Corr1. Warto jednak zwrócić uwagę, że ma więcej obserwacji odstających, które rekomendują obszary o rankingu powyżej 0.75. Pozytywny efekt wzrostu parametru p jest widoczny w przypadku testu Kołmogorowa-Smirnowa. Jego wzrost powoduje zarówno wybór obszarów o coraz lepszym rankingu ale również zmniejszenie jego różnorodności. Podobne wnioski dotyczą metody korzystającej ze zwykłej regresji logistycznej. Regiony otrzymane poprzez procedury Log1 i Log2 mają bardzo zbliżone wyniki. Obszary wybrane metodą Log2 mają nieco większą różnorodność rankingu i liczbę obserwacji odstających. Log2 wskazuje zarówno obszary o rankingu bliskim 1 jak i obszary o rankingu w granicach 0.75. Największa zmienność rankingu obszarów jest widoczna poprzez metodę Mix1 i pozostaje jednakowa przy zmianach parametru p. Metoda Mix2 radzi sobie podobnie jak metody Log1 i Log2. Jedynie dla $p = 0.02$ rekomenduje regiony o podobnej różnorodności jak Mix1. Procedury Wilc i Ttest wskazują również na podobne wnioski jak w przypadku regresji logistycznych. Niewątpliwie, rekomendacje za pomocą Wilc, Ttest, Log1, Log2, Mix2 podobnie jak w symulacji w części 1. są obiecujące i wymagają dalszej analizy. Interesujące wydają się również rezultaty otrzymane na podstawie testu Kołmogorowa-Smirnowa.



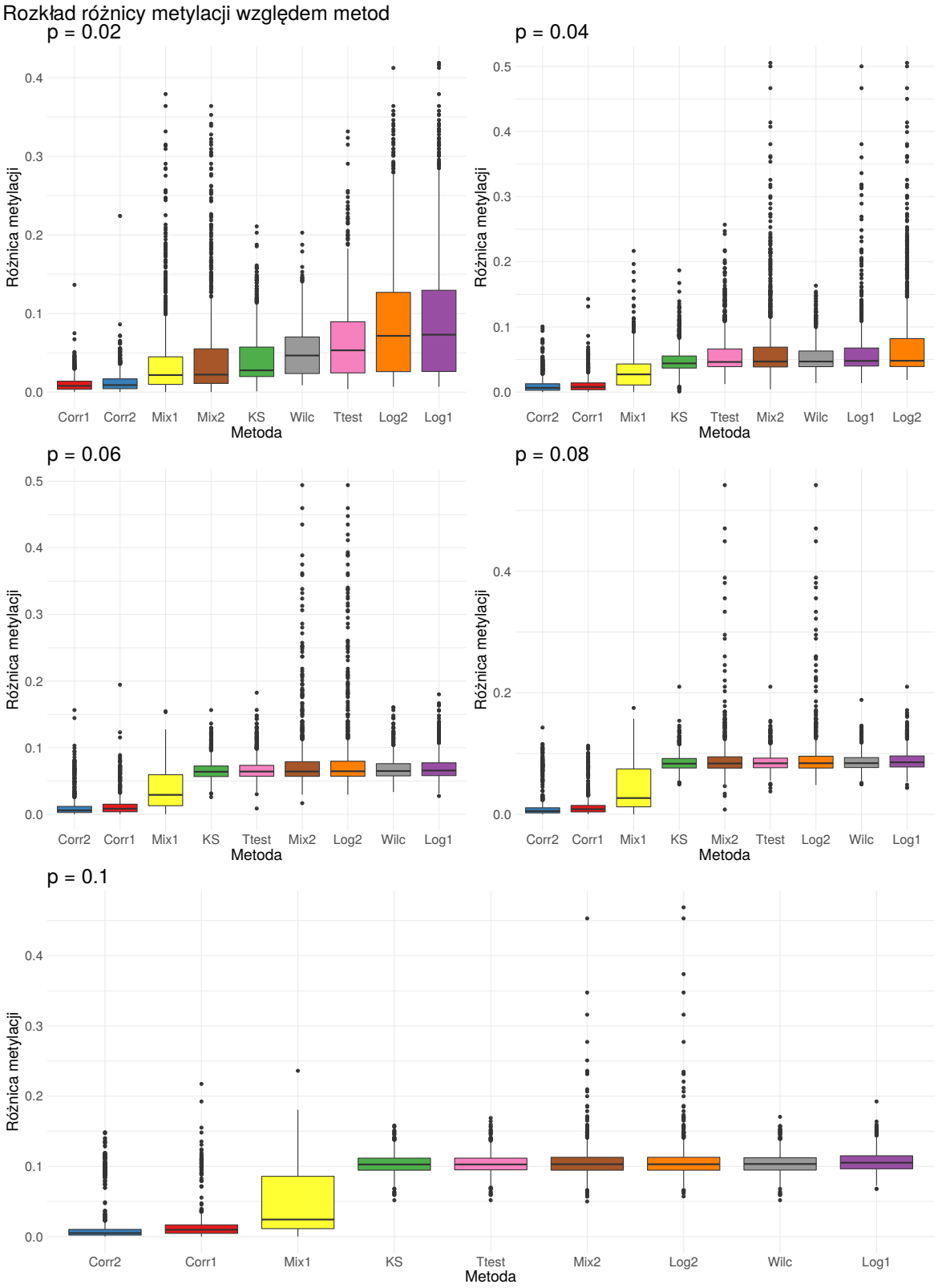
Rysunek 3.8: Rozkład rankingu top1000 obszarów w zależności od parametru p. Na osiach X przedstawiono uszeregowane rosnąco względem mediany kwantyli badane metody, na osi Y zaś rozkłady kwantyli za pomocą box-plotów.



Rysunek 3.9: Rozkład pokrycia top1000 obszarów w zależności od parametru p. Na osiach X przedstawiono uszeregowane rosnąco względem mediany pokrycia badane metody, na osi Y zaś rozkłady pokryć za pomocą box-plotów.

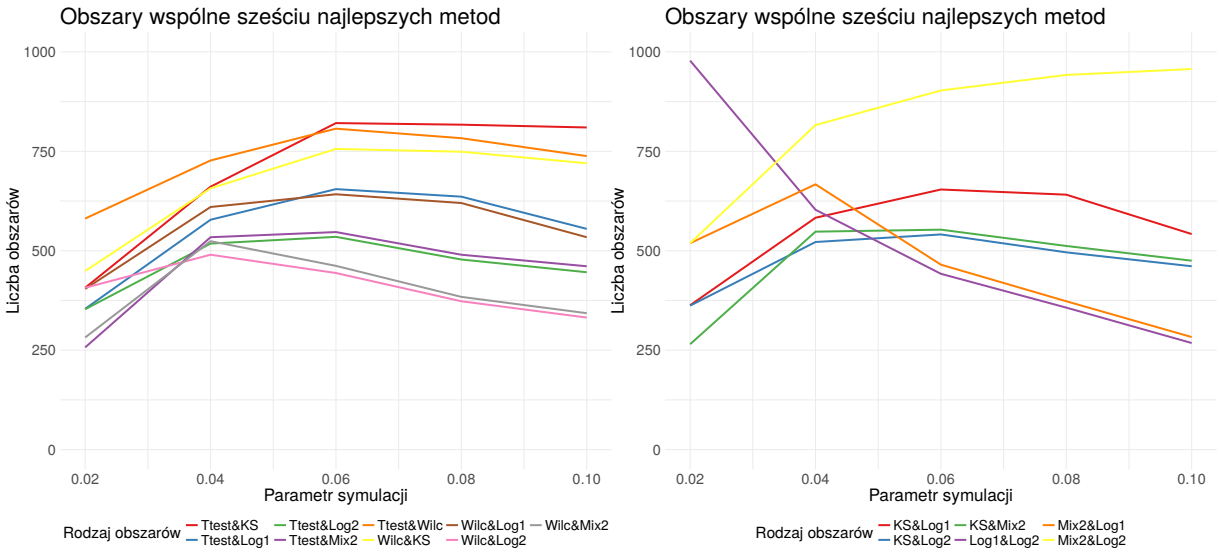
Rysunek 3.9 przedstawia rozkład pokrycia względem metod i dodatkowego parametru symulacji. Przy wzroście parametru p coraz bardziej klaruje się sytuacja różnorodności pokrycia regionów wybranych przez poszczególne sposoby. Obszary o najmniejszym pokryciu są rekomendowane przez metodę Mix1 i wahają się w graniach 8-28 obserwacji. Podobnym, lecz nieco większym pokryciem charakteryzują się regiony rekomendowane przez Log2 i Mix2. Metody Log i Corr rekomendują regiony o wyższym pokryciu w przypadku porównania wielkości efektu zmiennej grupującej niż wykorzystania jedynie wartości krytycznych. Dużą różnicę można zauważyć porównując Corr2 i Corr1. Corr1 wskazuje na obszary charakteryzujące się największą liczbą obserwacji w gronie wszystkich metod. Są to regiony o pokryciu w granicach 60 jednostek. Duże podobieństwo w pokryciu regionów można dostrzec konfrontując mechanizmy działania testów Wilcoxona, t-Studenta i Kołmogorowa-Smirnowa. Wszystkie wspomniane testy rekomendują regiony o liczbie obserwacji w granicach 15-36. Wzrost parametru p a więc różnicy średnich metylacji powoduje wybór obszarów o minimalnie większym pokryciu.

Rysunek 3.10 przedstawia natomiast rozkład różnicy stopnia metylacji między próbami wśród 1000 najlepszych obszarów. Największą różnorodność różnicy metylacji można zaobserwować w sytuacji, gdy parametr symulacji był niewielki i wynosił $p = 0.2$ lub $p = 0.4$. W przypadku niewielkiej przewagi ilości zmetylowanych nad niezmetylowanymi cytozynami w obu próbach najlepiej radzą sobie metody opierające się o standardową regresję logistyczną rekomendując obszary o stosunkowo największej różnicy metylacji. W miarę wzrostu parametru p metody Ks, Ttest, Wilc, Mix2, Log1, Log2 stają się coraz bardziej podobne pod względem rekomendowania obszarów o zbliżonej różnicy stopnia metylacji. Wybór obszarów, w których symulacja przebiegała z dodatkowym parametrem p był niezależny od stopnia pokrycia. Duża różnica stopnia metylacji była zatem możliwa zarówno dla obszarów o niewielkim jak i dużym pokryciu. Stąd, jeżeli jakaś metoda miała tendencję do wyboru obszarów o zadanym pokryciu, mogła wybrać obszary o dużych różnicach metylacji. Dlatego, odwrotnie jak w symulacji w części I trudno dostrzec różnorodność wyboru regionów względem metod na rysunku 3.10 wraz ze wzrostem p.



Rysunek 3.10: Rozkład różnicy metylacji top1000 obszarów w zależności od parametru p. Na osiach X przedstawiono uszeregowane rosnąco względem mediany różnicy metylacji badane metody, na osi Y zaś rozkłady różnic metylacji za pomocą box-plotów.

Wykresy 3.11 przedstawiają obszary wspólne metod, które wskazywały na najbardziej interesujące regiony na podstawie symulacji w części II.



(a) Liczba obszarów wspólnych

(b) Liczba obszarów wspólnych

Rysunek 3.11: Liczba obszarów wspólnych najlepszych metod w symulacji 2.

Wszystkie porównywane metody mają dosyć duży odsetek regionów, które polecają wspólnie. W przypadku przedstawionych wyników jest to wynik zawsze powyżej 25% dla top1000. Częstą sytuacją jest rekomendacja 40 - 60 % identycznych regionów, niezależnie od parametru symulacji czy porównywanej pary metod. Największe podobieństwo między polecanymi obszarami można znaleźć w metodach Mix2 i Log2, czy między testem t-Studenta i Kołmogorowa-Smirnowa oraz t-Studenta i Wilcoxona.

Miara, dzięki której łatwo można porównać wyniki między metodami to spodziewany odsetek wyników fałszywie dodatnich (FDR). Została ona obliczona następująco:

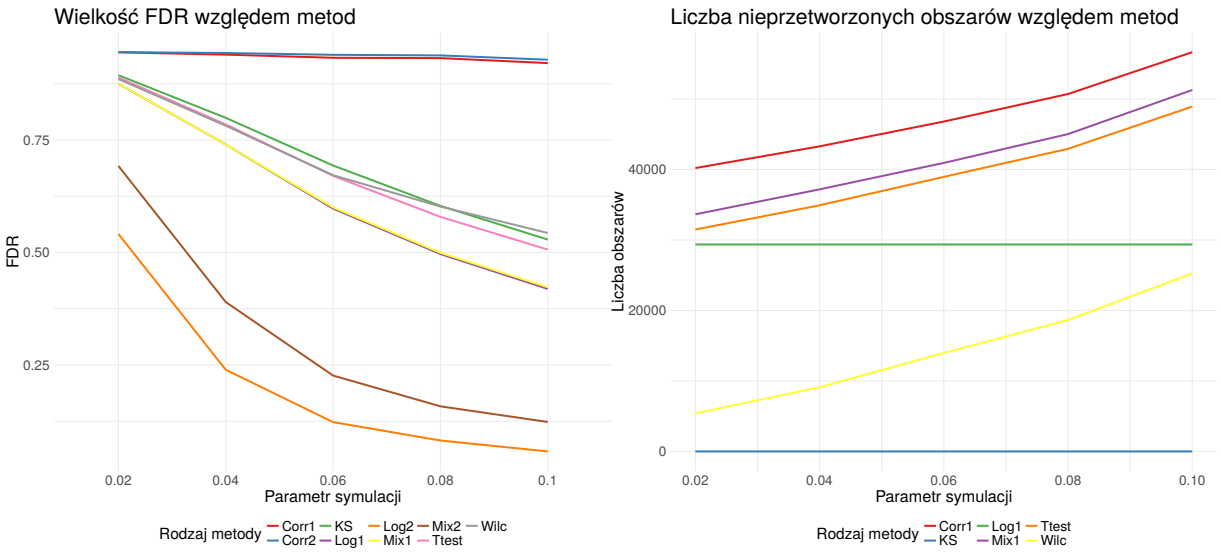
1. W każdej z metod uszeregowane zostały obszary pod względem atrakcyjności na podstawie mechanizmów, którymi posługiwano się do tej pory.

2. Następnie wybrano 5 % obszarów, które dana metoda uznała za najbardziej interesujące.

3. Obliczono iloraz obszarów znajdujących się w pierwszych 5 % regionów, dla których parametr symulacji stanowił średni poziom metylacji w obu grupach do ilości obszarów znajdujących się w pierwszych 5 % obszarów.

5 % regionów to liczba regionów, która posiadała prawdopodobieństwo sukcesu w przeprowadzonych symulacjach równe sumie średniego poziomu symulacji w danym regionie i parametru p. Miarę FDR można zatem utożsamiać z odsetkiem błędnie uznanych za interesujące obszarów,

w których nie powinna występować różnorodność metylacji. Na przytoczonym wykresie 3.12a można dostrzec trend wskazujący na zmniejszanie FDR wraz ze wzrostem p. Jedynie metody Corr utrzymują bardzo wysoki poziom miary niezmiennie od wskazanej symulacji. Duże wartości FDR dla $p = 0.02$ nie powinny zaskakiwać. Jest to niewielka różnica w prawdopodobieństwie sukcesu, która może nie prowadzić do zwiększenia ilości zasymulowanych ilości cytozyn, które uległy symulacji. Wynika to z tego, że pokrycie większości regionów jest niewielkie, a więc liczba prób Bernoulliego również. Zatem symulacje: $Bern(\overline{meth_{tkj}} + \overline{unmeth_{tkj}}, meth.rate_r)$ i $Bern(\overline{meth_{tkj}} + \overline{unmeth_{tkj}}, meth.rate_r + p)$ mogą prowadzić do zbliżonych wyników dla małej wartości p (podobnej liczby sukcesów). Zgonie z powyższym analizowane metody mogą mieć problem z wybraniem właściwych regionów. Procedury, dla których miara FDR wykazuje najkorzystniejsze wyniki to Log2 i Mix2. Są to metody, które posiłkują się wielkością efektu zmiennej grupującej a nie jedynie wartością krytyczną.



(a) Miara FDR względem metod

(b) Obszary nieprzetworzone jako [%] ogółu

Rysunek 3.12: Wykresy porównujące metody w symalacji II. Po lewej wykres miara FDR: na osi X zaznaczono kolejne symulacje, na osi Y miarę FDR. Po prawej liczba nieprzetworzonych obszarów: na osi X zaznaczono kolejne symulacje, na osi Y liczę obszarów. Na obu wykresach linie wskazują kolejne metody.

Również w tej metodzie symulacji warto zwrócić uwagę na ilość obszarów, które nie zostały przetworzone. Zgodnie z wykresem 3.12b, ilość regionów dla których nie uzyskano wyniku jest stała i niezależna od parametru p. Ponownie, wszystkie wyniki zostały przetworzone jedynie za pomocą testu Kołmogorowa-Smirnowa. Test Wilcoxona i regresja logistyczna mają niewielki odsetek liczby obszarów, dla których nie uzyskano rezultatów.

4. Opis pakietu metR

Pakiet można zainstalować korzystając z repozytorium znajdującym się na stronie `github.com`:

```
devtools::install_github('geneticsMiNIng/metR')
library(metR)
```

Analizę danych z badania metylacji należy rozpocząć od wczytania danych i przygotowania ich do żądanego formatu. Dalsze przetwarzanie będzie możliwe za pomocą funkcji *preprocessing*. Poniżej przykład wywołania funkcji bazujący na przykładowych danych zawartych w pakiecie *metR*. Zbiór *schizophrenia* zawiera dane pobrane ze strony: `http://www.neuroepigenomics.org/methylomedb/download.html`. Obserwacje dotyczące przebiegu metylacji zostały zsumowane na tej samej pozycji i chromosomie wśród grupy kontrolnej i grupy poddanej chorobie.

```
data('schizophrenia')
control <- schizophrenia %>% filter(category == 'control') %>%
dplyr::select(-category)
disease <- schizophrenia %>% filter(category == 'disease') %>%
dplyr::select(-category)
data <- preprocessing(control, disease)
head(data_all,4)
```

```
# A tibble: 4 x 7
#   chr      poz      prob     no     meth   unmeth   meth.rate
#   <chr>    <int>    <chr>  <int>  <int>  <int>    <dbl>
# 1      chr1   81412      x   35      29      6          0.829
# 2      chr1   81412      y   76      66      10         0.868
# 3      chr1   81442      x   35      29      6          0.829
# 4      chr1   81442      y   76      66      10         0.868
```

Wynikiem działania powyższej funkcji jest ramka danych, która stanowi część wspólną wyników dostępnych w dwóch podanych próbach na podstawie jednakowego chromosomu i pozycji.

Wybór wspólnych rezultatów wynika z faktu, że w celu zastosowania metod statystycznych na poszczególnych regionach konieczne są obserwacje w obu podpróbach dla wybranych chromosomów i pozycji.

Wynikowa ramka danych składa się z 7 następujących kolumn:

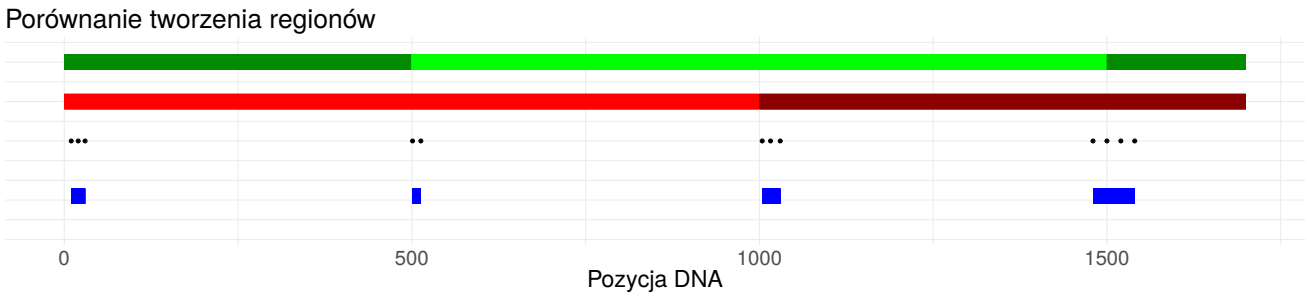
- *chr* - nazwa chromosomu
- *poz* - pozycja na chromosomie
- *prob* - nazwa próbki
- *no* - liczba wszystkich sekwencji
- *meth* - liczba sekwencji, dla których odbył się proces metylacji
- *unmeth* - liczba sekwencji, dla których nie odbył się proces metylacji
- *meth.rate* - stopień metylacji rozumiany poprzez stosunek wartości *meth/no*

Argumentami funkcji *preprocessing* są dwie ramki danych, które powinny zawierać wyżej wymienione kolumny pomijając kolumnę *prob*.

Następnym etapem analizy danych z badania metylacji jest utworzenie regionów, dla których przeprowadzone zostaną osobne testy weryfikujące istnienie różnicy stopnia metylacji w dwóch próbach. Możliwe jest wykorzystanie dwóch sposób kreacji grup testowych:

1. funkcja *create_tiles_max_gap(data = data, gaps.length = 100)*, która tworząc regiony wykorzystuje obserwacje, których pozycja na chromosomie jest oddalona maksymalnie o argument *gaps.length*. Wynikiem działania funkcji będzie ramka danych *data* podana jako argument z dodatkową kolumną *tiles* wskazującą na przynależność danej obserwacji do regionu.

2. funkcja *create_tiles_fixed_length(data = data, tiles.length = 2000, common = T)*, która przechodzi oknem ruchomym o długości równej *tiles.length* z krokiem równym *tiles.length* gdy argument *common* = *F* lub z krokiem *tiles.length/2* gdy *common* = *T* i wykorzystuje obserwacje znajdujące się w danym przedziale jako jeden region. Wynikiem działania funkcji będzie ramka danych *data* podana jako argument z dodatkową kolumną *tiles* wskazującą na przynależność danej obserwacji do regionu. Gdy *common* = *T* ramka danych będzie posiadała kolejną kolumnę *tiles.common*, która wskazuje na przynależność danej obserwacji do dodatkowych regionów.



Rysunek 4.1: Porównanie działania funkcji *create_tiles_max_gap* oraz *create_tiles_fixed_length*. Niebieskie linie oznaczają obszary wygenerowane za pomocą *create_tiles_max.gap*, czerwone linie to obszary wygenerowane za pomocą *create_tiles_fixed_length* z argumentem *common* = *F* a zielone linie to dodatkowe regiony, gdy *common* = *T*.

Wywołując funkcję *create_tiles_max_gap(data = data, gaps.length = 20)* dla przedstawionych powyżej pozycji otrzymamy 4 regiony z następującymi zbiorami pozycji: $\{10, 20, 30\}$, $\{501, 513\}$, $\{1004, 1016, 1030\}$, $\{1480, 1500, 1520, 1540\}$. Korzystając z *create_tiles_fixed_length(data = data, tiles.length = 1000, common = F)* otrzymamy 2 regiony z następującymi zbiorami pozycji: $\{10, 20, 30, 501, 513\}$, $\{1004, 1016, 1030, 1480, 1500, 1520, 1540\}$. Korzystając z powyższej funkcji z argumentem *common* = *T* otrzymamy jeszcze regiony ze zbiorami: $\{10, 20, 30\}$, $\{501, 513, 1004, 1016, 1030, 1480\}$, $\{1500, 1520, 1540\}$.

Po utworzeniu regionów warto sprawdzić podstawowe statystyki. Do tego można wykorzystać funkcję *get_stats*.

```
data.tiles <- create_tiles_max_gap(data = data, gaps.length = 100)
stats <- get_stats(data.tiles)
```

Uzyskana ramka danych składa się z 14 następujących kolumn:

- *chr* - nazwa chromosomu
- *start* - pierwsza pozycja regionu na chromosomie
- *end* - ostatnia pozycja regionu na chromosomie
- *meth.cov* - liczba obserwacji w zadanym regionie
- $meth.max_x, meth.mean_x, meth.min_x, meth.sd_x$ maksymalna, średnia, minimalna wartość oraz odchylenie standardowe stopnia metylacji w 1. podanej próbce
- $meth.max_y, meth.mean_y, meth.min_y, meth.sd_y$ maksymalna, średnia, minimalna wartość oraz odchylenie standardowe stopnia metylacji w 2. podanej próbce

- *meth.diff* - wartość bezwzględna różnicy stopnia metylacji w dwóch próbach
- *quantile* - poziom kwantyla bezwzględnej różnicy stopnia metylacji estymowany na podstawie regresji kwantylowej uwzględniającej liczbę obserwacji w regionie

Najbardziej istotną funkcją dostępną w *metR* jest *fin_DMR*.

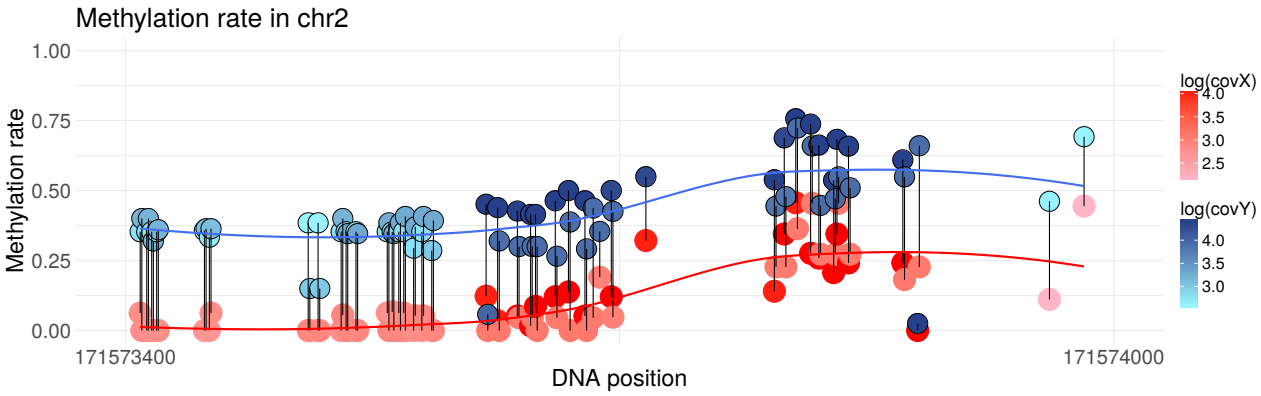
```
result <- find_DMR(data = data.tiles,
        methods = c('Wilcoxon', 'Ttest', 'KS', 'Reg.Log','Reg.Mixed',
        'Reg.Corr.Mixed'),  p.value.log.reg = 0.01,
        p.value.reg.corr.mixed = 0.01, p.value.reg.mixed = 0.05)
```

```
head(result$Wilcoxon,4)
# A tibble: 4 x 4
# chr     start       end     p.value
# <chr>     <int>     <int>       <dbl>
# 1  chrM       106      1026 0.00139423
# 2 chr20     64381     64518 0.01153999
# 3  chr4     10674     10934 0.01725576
# 4  chr2     14188     14238 0.02130318
```

Dzięki niej możliwe jest posortowanie regionów względem najbardziej interesujących różnic stopnia metylacji. Dostępne metody to: testy: Wilcoxona, t-Studenta, Kołmogorowa-Smirnowa oraz regresje logistyczna, regresja logistyczna z efektami losowymi oraz regresja logistyczna z efektami losowymi uwzględniająca korelację. Szczegółowy opis poszczególnych metod został już przedstawiony w rozdziale 1. i 2. niniejszej pracy. Wynik działania *find_DMR* to lista ramek danych, których nazwy odpowiadają nazwie metody wskazanej w wektorze *methods*. Poszczególna ramka danych zawiera informację o minimalnej i maksymalnej pozycji w grupie, chromosomie oraz o wartości krytycznej przeprowadzonego testu i ewentualnie parametrze oszacowania zmiennej grupującej. Wyniki na podstawie testów statystycznych są posortowane rosnąco względem wartości krytycznych. Wyniki regresji logistycznych są posortowane malejąco względem wartości bezwzględnej oszacowania zmiennej grupującej wśród obserwacji, których wartość krytyczna zmiennej grupującej jest mniejsza od wskazanego parametru (*p.value.log.reg, p.value.reg.corr.mixed, p.value.reg.mixed* w zależności od metody.) W przypadku gdy nie zostanie podana wartość parametru potrzebnego do wybrania regionów, obszary zostają posortowane rosnąco względem wartości krytycznej określającej istotność statystyczną parametru dla zmiennej grupującej.

W celach prezentacji graficznej badanych obszarów może posłużyć funkcja *draw_methylation*. Należy podać ramkę danych (argument *data*) na podstawie której zostaną pobrane parametry dla regionu. Może to być ramka danych będąca wynikiem funkcji *preprocessing* lub *create_tiles_max_gap* czy *create_tiles_fixed_length*. Następnie wystarczy podać minimalną i maksymalną pozycję oraz chromosom regionu, który chcemy narysować.

```
draw_methylation(data = data, start = 171573250,
end = 171574000, chr = 'chr2')
```



Rysunek 4.2: Wynik działania funkcji *draw_methylation*

Wykres 4.2 prezentuje stopień metylacji w zadanym regionie w dwóch próbach. Dodatkowo kolorem została zaznaczona ilość sekwencji dla każdej obserwacji na skali logarytmicznej.

Bardziej szczegółowy opis pakietu *metR* oraz przykład kompletnej analizy zastosowanej do danych zamieszczonych w pakiecie można znaleźć na stronie repozytorium pakietu `https://github.com/geneticsMiNIng/metR` przeglądając dokumentację oraz winietkę pakietu.

5. Podsumowanie

Praca przedstawia porównanie różnych algorytmów mających na celu wskazanie obszarów o dużej różnicy w stopniu metylacji w dwóch grupach. Zostało również opisane podłoże merytoryczne potrzebne do analizy danych z badania metylacji. Skorzystano także z gotowych implementacji zbadanych metod i dostosowano je do postaci problemu. Efektem wyżej wymienionych prac jest pakiet *metR*. Umożliwia on kreację obszarów, dla których przeprowadzone zostaną testy statystyczne potrzebne do wskazania obszarów o największej pod względem różnicy metylacji atrakcyjności. Następnie uszeregowanie tak uzyskanych regionów na podstawie dostępnych algorytmów od najbardziej do najmniej interesującego. Pakiet umożliwia również prezentację graficzną obszarów. W pracy opisano podstawy matematyczne metod, które zostały już wykorzystane w analizie danych metylacyjnych jak również metody, które nie były do tej pory szeroko wykorzystywane. Do algorytmów, z których korzystano najczęściej należą test medianowy Wilcoxona oraz test średnich t-Studenta. Mniej popularne metody to skorzystanie z regresji logistycznej oraz z regresji logistycznej z efektami losowymi czy testu Kołmogorowa-Smirnowa.

Na podstawie przeprowadzonych symulacji można zauważyć, że rzadziej używane algorytmy radzą sobie równie dobrze jak te bardziej rozpowszechnione. Na szczególne uznanie zasługuje posłużenie się standardową regresją logistyczną. Wykorzystano podejście, w którym obszary są uszeregowane względem wartości krytycznej zmiennej grupującej próby lub względem wielkości efektu zmiennej grupującej. Pierwszy sposób rekomenduje obszary o największej bezwzględnej różnicy poziomu metylacji w dwóch grupach. Taki wynik otrzymano, w sytuacji odtworzenia rzeczywistego charakteru danych w symulacji w części I.

Drugi sposób sugeruje natomiast najwięcej regionów o faktycznej różnicy metylacji. Podobnie radzi sobie regresja logistyczna z efektami losowymi. Przedstawiony rezultat uzyskano w warunkach sztucznego wygenerowania obszarów wśród których część miała dodatkowy parametr symulacji odpowiedzialny za zwiększenie różnorodności między grupami. Regresja logistyczna z efektami losowymi ma znacznie zwiększony, nawet kilkunastokrotnie czas przetwarzania danych w porównaniu ze zwykłą regresją logistyczną. Ze względu na wysokowymiarową postać danych, lepiej zatem skorzystać z prostszej formy, ponieważ rezultaty obu podejść są zbliżone.

Na uznanie zasługuje test Kołmogorowa-Smirnowa, który jako jedyny jest w stanie zwró-

cić wartość krytyczną nawet, gdy ma niewielką liczbę obserwacji w regionie. Kwestią sporną pozostaje sensowność obliczonej miary w takich sytuacjach.

Każdy z algorytmów charakteryzuje się rekomendacją obszarów o nieco odmiennej ilości obserwacji. Ciekawym wynikiem byłoby więc nie sugerowanie się nie tylko jedną metodą. Można byłoby więc w takim przypadku otrzymać obszary o stosunkowo dużej ilości obserwacji ale mniejszej różnicy metylacji za pomocą testu Wilcoxona i największych różnicach metylacji ale mniejszej ilości obserwacji za pomocą jednego ze sposobów związanych z regresją logistyczną.

Bibliografia

[1] Allaire J. et al. *rmarkdown: Dynamic Documents for R*, `https://CRAN.R-project.org/package=rmarkdown`, 2017.

[2] Aryee M.J. et al. *Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays*, Bioinformatics 30(10), pp.1363-1369, 2014.

[3] Assenov Y. et al. *Comprehensive Analysis od DNA Methylation Data with RnBeads*, Nature Methods, 11(11):1138-1140, 2014.

[4] Bache S., Wickham H. *magrittr: A Forward-Pipe Operator for R* `https://CRAN.R-project.org/package=magrittr`, 2014.

[5] Barfield R. et al. *CpGassoc: Association Between Methylation and a Phentype of Interest*, 2017.

[6] Bates D. et al. *Fitting Linear Mixed-Effects Models Using lme4*,Journal of Statistical Software, 67(1), 2015.

[7] Blackman J. *An Extension of the Kolmogorov Distribution*,The Annals of Mathematical Statistics, Vol. 27, No. 2, 1956.

[8] Breslow E., Clayton D. *Approximate Inference in Generalized Linear Mixed Models*,Journal of the America Statistical Association, Vol. 88. No. 421, 1993.

[9] Chen J. et al. *A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution*, Biostatistics, 3,3, pp.347-360, 2002.

[10] Czepiel S. *A Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*, `http://czep.net/stat/mlelr.pdf`, 2015.

[11] Dowle M., Srinivasan A., *data.table: Extension of 'data.frame'*, `https://CRAN.R-project.org/package=data.table`, 2017.

[12] Du P., Bourgon R. *methyAnalysis: DNA methylation data analysis and visualization*, 2017.

[13] Harville D. *Maximum Likelihood Approaches to Variance Components Estimation and to Related Problems*, Journal of the American Statistical Association, Vol. 72. No. 358, pp. 320-338, 1977.

[14] Jaffe A. et al. *Bump hunting to indentify differentially methylated regions in epigenetic epodemiology studies*, International Journal of Epidemiology, 41:200-209, 2012.

[15] Jang W., Lim J. *A Numerical Study of PQL Estimation Biases in Generalized Linear Mixed Models Under Heterogeneity of Random Effects*,Communications in Statistics - Simulation and Computation, 38: 692-702, 2009.

[16] Koenker R., *Quantile regression*, J. of Economics Perspectives, 2001.

[17] Koenker R., Basset G., *Regression Quantiles*, Econometrica 46 No.1, 33-50, 1978.

[18] Li D. et al. *An evaluation of statistical methods for DNA microarray data analysis*, BMC Bioinformatics, 16-217, 2015.

[19] Łukasik M. et al. *Wpływ metylacji DNA na funkcjonowanie genomu*, Warszawski Uniwersytet Medyczny `http://biuletynfarmacji.wum.edu.pl/0902Lukasik/Lukasik.pdf`, 2009.

[20] Maksimovic J. et al. *SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips*, Genome Biology 13-R44, 2012.

[21] Pinheiro J. et al. *nlme: Linear and Nonlinear Mixed Effects Models*, 2017.

[22] Smyth G. K. *Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments*, Statistical Applications in Genetics and Molecular Biology Volume 3, Issue 1, 2004.

[23] Topolski K. *Notatki do przedmiotu Statystyka*, Uniwersytet Wrocławski, `http://www.math.uni.wroc.pl/~topolski/STAT.htm`, 2017.

[24] Wang D. et al. *IMA: an R package for high-through analysis of Ilumina's 450K Infiunium methylation data*, Bioinfmatics 28(5) pp.729-30., 2017.

[25] Wickham H. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, `http://ggplot2.org`, 2009.

[26] Wickham H. et al. *roxygen2: In-Line Documentation for R.*, `https://CRAN.R-project.org/package=roxygen2`, 2017.

[27] Wickham H. et al. *readr: Read Rectangular Text Data*, `https://CRAN.R-project.org/package=readr`, 2017.

[28] Wickham H., Lionel H. *Easily Tidy Data with 'spread()' and 'gather()' Functions*, `https://CRAN.R-project.org/package=tidyr`, 2017.

[29] Wickham H. et al. *dplyr: A Grammar of Data Manipulation*, `https://CRAN.R-project.org/package=dplyr`, 2017.

[30] Venables W. et al. *Modern Applied Statistics with S.*, Fourth Edition. Springer, New York, `http://www.stats.ox.ac.uk/pub/MASS4`, 2002.

[31] Yadolah D. *The Concise Encyclopedia of Statistics*, Spinger Science + Business Media 2008.

Wykaz symboli i skrótów

bp	jednostka długości regionu DNA
Corr1	regresja logistyczna z efektami losowymi z uwzględnieniem autokorelacji z sortowaniem obszarów po wielkościach krytycznych
Corr2	regresja logistyczna z efektami losowymi z uwzględeniem autokorelacji z sortowaniem obszarów po wielkości efektu
DMR	ang. *Differentially methylated region* - Obszar o zróżnicowanym stopniu metylacji
FDR	ang. *False discovery rate* - Spodziewany odsetek wyników fałszywie dodatnich
KS	test Kołmogorowa-Smirnowa
Log1	regresja logistyczna z sortowaniem obszarów po wartościach krytycznych
Log2	regresja logistyczna z sortowaniem obszarów po wielkości efektu
meth	liczba cytozyn, które uległy metylacji
Mix1	regresja logistyczna z efektami losowymi z sortowaniem obszarów po wartościach krytycznych
Mix2	regresja logistyczna z efektami losowymi z sortowaniem obszarów po wielkości efektu
SWAN	ang. *Subset-quantile Within Array Normalization* - Metoda normalizacji danych
Ttest	test t-Studenta
unmeth	liczba cytozyn, które nie uległy metylacji
Wilc	test znaków Wilcoxona
wyspy CpG	regiony w genomie o podwyższonej zawartości cytozyn w stosunku do przeciętnej dla całego genomu

Spis rysunków

Spis tabel