Metody obliczeniowe w analizie różnicowej macierzy kontaktów chromatynowych

Rafał Zaborowski

Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

20 maja 2020

Autoreferat

---

1 Wstęp

Od zakończenia projektu poznania genomu ludzkiego (z ang. Human Genome Project) ilość danych biologicznych zbieranych w badaniach genomowych rośnie w sposób wykładniczy [Cook et al., 2015]. Aby nadążyć za analizą stale napływających danych, rośnie zapotrzebowanie na opracowywanie nowych metod obliczeniowych, które ułatwią przetwarzanie ogromnych ilości pozyskiwanych informacji.

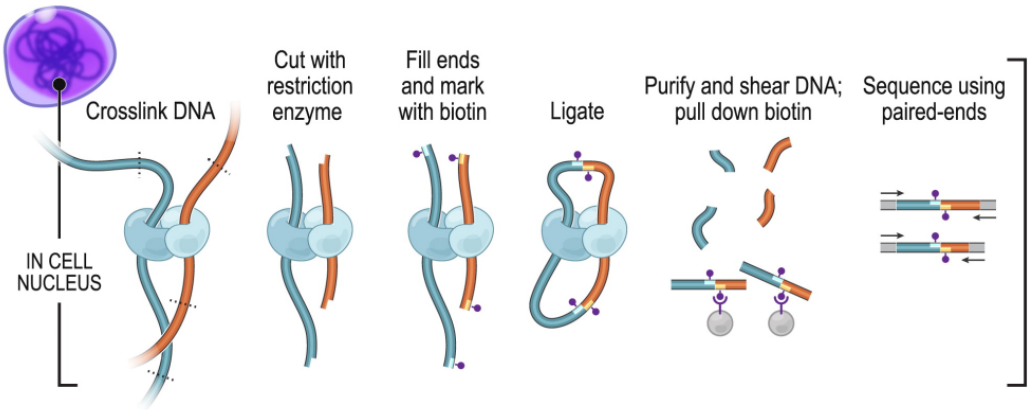
Jednym z najbardziej znanych przykładów złożonego problemu obliczeniowego, który jest powszechny we współczesnej genomice, jest mapowanie krótkich odczytów. Przywołany problem jest bardzo istotny, ponieważ wiele protokołów biologicznych jest opartych na technologiach sekwencjonowania nowej generacji (z ang. Next Generation Sequencing), wytwarzających miliony krótkich odczytów, których pozycja w genomie musi zostać określona, aby zinterpretować wyniki eksperymentu [Flicek and Birney, 2009]. Naiwne rozwiązania tego problemu są bezużyteczne ze względu na wysokie koszty obliczeniowe, a powszechnie stosowane podejścia wykorzystują zaawansowane algorytmy.

Innym przykładem jest problem pomiaru różnic w ekspresji genów mający ogromne znaczenie, między innymi w badaniach farmaceutycznych [Bai et al., 2013, Lohmann et al., 2013, Lee et al., 2017]. Pomiary ekspresji genów są zwykle przeprowadzane przez ilościowe określenie liczby odczytów mapowanych do docelowych sekwencji DNA. W tym kontekście zakłada się, że liczba zliczeń danej sekwencji w próbce jest proporcjonalna do ekspresji określonego genu. Jednak ze względu na liczne błędy towarzyszące eksperymentom sekwencjonowania nowej generacji bezpośrednie porównania zliczeń zwykle nie są przydatne, a badacze wykorzystują dedykowane modele statystyczne, które pozwalają na poprawne zinterpretowanie uzyskanych wyników [Froussios et al., 2019].

Ważnym zagadnieniem w genomice jest badanie mechanizmów różnicowania komórek. Dużo wysiłku poświęca się zrozumieniu przyczyn odpowiedzialnych za różnorodność komórkową organizmu zważywszy na to, że jądra komórkowe zawierają prawie identyczną sekwencję DNA. Jednym z wielu czynników, które mają istotny wpływ na różnicowanie i

funkcjonowanie komórek, jest przestrzenna organizacja chromatyny [Zheng and Xie, 2019]. W odniesieniu do badań strukturalnych jąder komórkowych ostatnia dekada była szczególnie owocna ze względu na postęp w technikach badania architektury chromosomów zwanych 3C (z ang. Chromosome Conformation Capture).

Niniejsza rozprawa jest poświęcona opracowaniu metod obliczeniowych do analizy danych uzyskanych z metody Hi-C - wysoko rozdzielczego wariantu protokołu 3C. W szczególności rozprawa koncentruje się na analizie porównawczej, której celem jest odkrywanie istotnych biologicznie różnic struktury chromatyny pomiędzy badanymi próbkami.



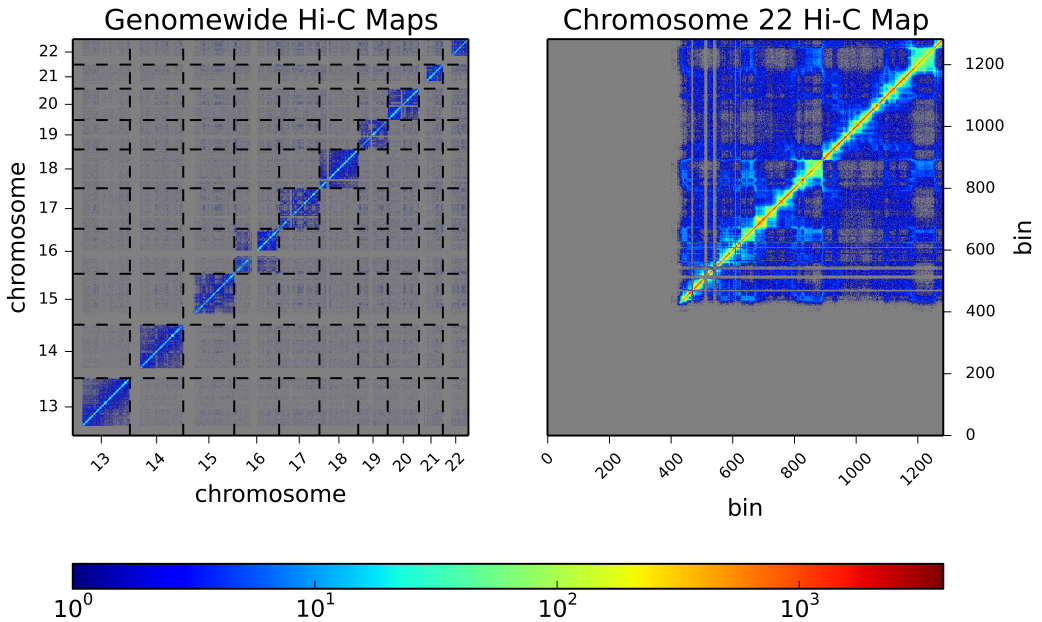
Rysunek 1. Schemat protokołu Hi-C. Rysunek z [Rao et al., 2014]. Użyto za zgodą Elsevier.

1.1 Badania nad strukturą chromatyny

Przez długi czas badania struktury genomu były ograniczone głównie z przyczyn technicznych. Pierwsze badania organizacji chromatyny zostały przeprowadzone przy użyciu mikroskopii świetlnej przez Carla Rabla, który zasugerował koncepcję modelu terytoriów chromosomowych (z ang. chromosome territories) w 1885 roku [Cremer and Cremer, 2010]. Jego hipoteza głosiła, że chromosomy zajmują odrębne objętości jądra komórkowego, zamiast przeplatać się nawzajem. Koncepcja terytoriów chromosomowych została ostatecznie potwierdzona przez Thomasa i Christopha Cremera w latach 80-tych [Cremer and Cremer, 2001].

Rozwój mikroskopii elektronowej umożliwił bardziej szczegółowe badania organizacji chromatyny, co pozwoliło na zaobserwowanie włókna chromatynowego w postaci "sznura koralików" (z ang. beads-on-string) [Annunziato, 2008]. Kolejnym postępem w badaniu architektury genomu był rozwój fluorescencyjnej hybrydyzacji in situ (FISH - z ang. Fluorescence In Situ Hybridization) [Gall and Pardue, 1969]. Ta technika pozwala ustalić, czy wybrane fragmenty włókna chromatyny znajdują się w bezpośrednim sąsiedztwie w przestrzeni. Pomimo dość żmudnego i kosztownego protokołu FISH okazał się niezwykle przydatny w wielu ważnych eksperymentach i nadal jest wykorzystywany do badania kolokalizacji elementów regulatorowych [Stack et al., 1977, Zorn et al., 1979].

Kamieniem milowym w badaniach struktury chromatyny było opracowanie technik 3C (ang. Chromosome Conformation Capture) [Dekker et al., 2002]. Ta grupa metod pozwala badać architekturę genomu w wysokiej rozdzielczości, poprzez próbkowanie milionów interakcji chromatynowych w trakcie pojedynczego eksperymentu. Najbardziej zaawansowany wariant techniki 3C - Hi-C umożliwia prowadzenie badań strukturalnych na niespotykaną dotąd



Rysunek 2. Mapy kontaktów Hi-C ludzkich komórek IMR90 z badania przeprowadzonego przez [Rao et al., 2014]. Lewa macierz ilustruje wewnątrz- i między-chromosomowe mapy kontaktów chromosomów 13-22. Granice chromosomów są oznaczone przerywanymi liniami pionowymi i poziomymi. Prawa macierz przedstawia wewnątrz-chromosomową mapę kontaktów chromosomu 22. Pasek kolorów wskazuje intensywność kontaktów (liczbę interakcji). Szare obszary oznaczają brak interakcji. Duże, szare paski obecne w większości macierzy odpowiadają centromerom, telomerom lub innym regionom niemapowalnym.

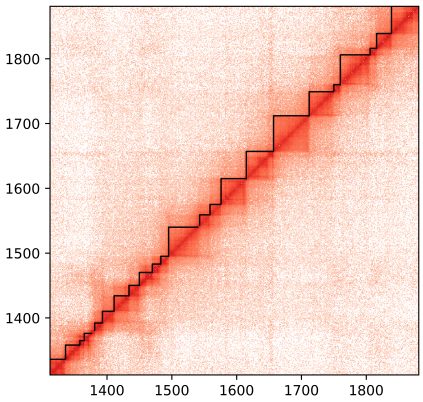
skalę [Lieberman-Aiden et al., 2009]. Podczas eksperymentu Hi-C pary fragmentów chromatyny, które znajdują się w przestrzennym sąsiedztwie, są stabilizowane przez kowalencyjne wiązanie formaldehydem w procesie zwanym sieciowaniem (z ang. crosslinking, Rysunek 1, jasnoniebieski obwarzanek). Następnie próbka jest trawiona enzymem restrykcyjnym, końce fragmentów restrykcyjnych są wypełniane biotyną i ligowane. Później DNA jest fragmentowany za pomocą sonikacji lub enzymu restrykcyjnego, a zligowane elementy są wyciągane z roztworu przy użyciu specyficznego wiązania biotyny ze streptawidyną. Na koniec uzyskane fragmenty sekwencjonuje się metodą sparowanych odczytów (z ang. paired-end) w celu uzyskania biblioteki odczytów, która wskazuje kontakty pomiędzy włóknem chromatynowym. Po sekwencjonowaniu odczyty są mapowane na genom referencyjny, aby określić ich pochodzenie i wskazać interakcje. Standardowym sposobem gromadzenia danych Hi-C są mapy kontaktów, które agregują kontakty między dowolną parą fragmentów DNA o stałej, z góry określonej długości (Rysunek 2).

Jednym z najbardziej charakterystycznych wzorów w mapach kontaktów Hi-C genomów ssaków jest istnienie kwadratowych bloków o wysokiej intensywności oddziaływań, które są zlokalizowane wzdłuż głównej przekątnej (Rysunek 3) [Sexton et al., 2012, Dixon et al., 2012]. Obecność tych jednostek, nazwanych później TAD-ami (z ang. Topologically Associating Domains), została potwierdzona w wielu organizmach [Dixon et al., 2012, Rudan et al., 2015]. Zainteresowanie badaniem TAD-ów i ich korelacji z ekspresją genów lub epigene-

tycznymi modyfikacjami chromatyny zaowocowało opracowaniem licznych algorytmów do systematycznego wykrywania tych obiektów [Zufferey et al., 2018].

1.2 Związek pomiędzy strukturą chromatyny a regulacją genów

Większość sekwencji ludzkiego DNA składa się z regionów niekodujących. Pewna klasa takich regionów zwana elementami regulatorowymi odgrywa kluczową rolę w utrzymywaniu procesów komórkowych [Maston et al., 2006]. Zasadniczo elementy regulatorowe można podzielić na dwie rodzaje: bliskie i odległe (Rysunek 4a). Te pierwsze zawierają motywy rozpoznawane przez polimerazę DNA, czynniki transkrypcyjne i elementy odległe. Zazwyczaj transkrypcja specyficzna tkankowo jest wyzwalana po tym, jak fragment DNA zwany enhancerem (z ang. enhancer - wzmacniacz transkrypcji) zwiąże się z regionem promotorowym genu (Rysunek 4b). W przeciwieństwie do enhancerów, wyciszacze (z ang. silencers) związane z promotorem działają jako represory danego genu. Izolatory z kolei (z ang. insulators) ograniczają zasięg elementów regulatorowych (Rysunek 4). Izolują one różne obszary genomu od niepożądanego wzajemnego wpływu na ich regulację (Rysunek 4a).



Rysunek 3. Fragment mapy kontaktów Hi-C chromosomu 18 ludzkich embrionalnych komórek macierzystych (ESC) z nałożonymi TAD-ami (czarne trójkąty wzdłuż głównej przekątnej). Dane Hi-C z pracy [Dixon et al., 2012]

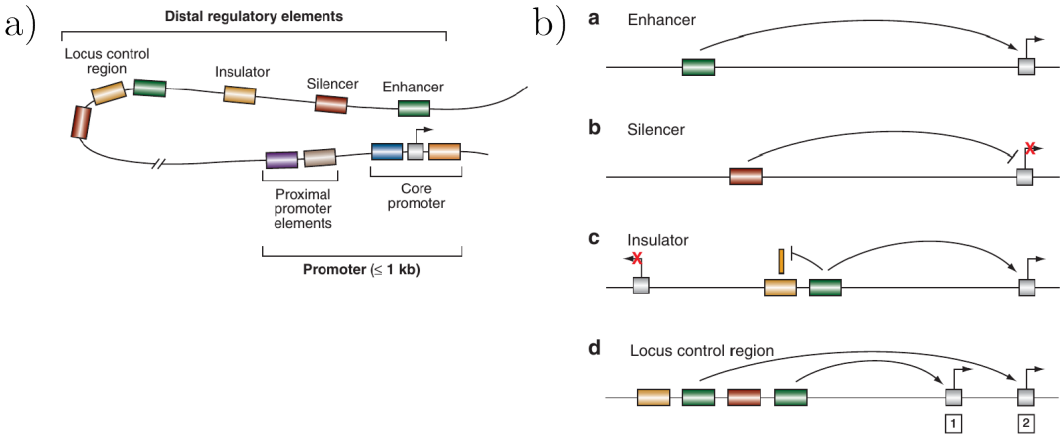
Funkcjonowanie dystalnych elementów regulatorowych wymaga tworzenia oddziaływań dalekiego zasięgu między odległymi fragmentami włókna chromatynowego. Ostatnie dniesienia z analiz danych Hi-C sugerują istotną rolę TAD-ów w kształtowaniu krajobrazów regulacyjnych poprzez ograniczanie zasięgu elementów regulatorowych. W pracy [Andrey et al., 2013] wykazano, że zaburzenie segmentacji domenowej chromosomu może prowadzić do poważnych wad rozwojowych i w rezutacie do chorobowego fenotypu. Powyższe obserwacje podkreślają znaczenie badań architektury genomu w pracach nad wyjaśnianiem mechanizmów regulacji genów.

2 Główne wyniki

Badania Hi-C mogą przyczynić się do znacznej poprawy naszego zrozumienia organizacji genomów. Jednak protokół ten jest bardzo złożony, a przetwarzanie uzyskanych danych jest czasochłonne i wymaga skrupulatnej analizy statystycznej aby oddzielić interesujący badaczy sygnał od szumu. Błędy powstające w trakcie protokołu Hi-C utrudniają bezpośrednie porównania między różnymi zbiorami danych przez co niezbędnym staje się zastosowanie technik statystycznych.

W tej pracy badaliśmy problem porównywania zbiorów danych Hi-C na poziomie domen topologicznych, a także par fragmentów DNA. W pierwszym przypadku zaproponowaliśmy nową miarę odległości dostosowaną do porównywania segmentacji TAD. W drugim przypadku

opracowaliśmy podejście do wykrywania interakcji różnicowych macierzy kontaktów Hi-C w oparciu o uogólnione modele liniowe.



Rysunek 4. Elementy regulatorowe. a) Region regulatorowy genów zawiera promotor oraz elementy odległe, takie jak izolatory, wyciszacze i enhancery. Regulacja niektórych genów jest inicjowana poprzez tworzenie oddziaływań dalekiego zasięgu pomiędzy enhancerami i promotorami. b) Funkcje odległych elementów regulatorowych. Enhancery inicjują transkrypcję poprzez interakcje dalekiego zasięgu. W podobny sposób wyciszacze zapobiegają transkrypcji genów. Izolatory ograniczają zasięg enhancerów i wyciszaczy. Odległe elementy regulatorowe są zwykle zorganizowane w tzw. regiony kontroli locus (z ang. locus control region). Rysunek z [Maston et al., 2006]. Użyto za zgodą Annual Reviews Inc.

2.1 Porównywanie segmentacji chromosomów

W tej części wprowadzamy nową miarę odległości do porównań segmentacji chromosomów zwaną miarą BP. Problem porównywania segmentacji chromosomów jest istotny w badaniach nad architekturą genomu podczas:

- klasyfikacja typów komórek według podobieństwa organizacji struktury chromatyny,
- oceny jakości replikacji eksperymentu Hi-C,
- porównywania algorytmów do wyznaczania TAD-ów.

Zaczynamy od ustalenia notacji. Formalnie segmentacja chromosomu jest zbiorem $X \subseteq \{i|\ i \in \mathbb{Z}, 0 < i < N\}$, który odpowiada podziałowi odcinka długości N na zbiór segmentów (Rysunek 5a).

Definicja 1 *Segmentem nazywamy pół-otwarty, niepusty zbiór dyskretny:* $(a, b] = \{x \in \mathbb{N}^+ \mid a < x \leq b\}$ *z następującymi relacjami:*

(i) równość: $(a, b] = (c, d] \iff a = c \wedge b = d$

(ii) zawieranie: $(a, b] \subseteq (c, d] \iff a \geq c \wedge b \leq d$

(iii) koniunkcja:

$$(a,b] \cap (c,d] \iff \begin{cases} \varnothing, & \textit{gdy } a \geq d \vee c \geq b \\ (\max(a,c), \min(b,d)], & \textit{w p. p.} \end{cases}$$

Segmentami mogą być chromosomy, TAD-y oraz ich części wspólne. Każdy chromosom może zostać podzielony na zbiór (uporządkowany) nie nakładających się, kolejnych segmentów - TAD-ów lub regionów nie będących TAD-ami. W dalszej części tekstu dla uproszczenia ograniczamy nasze rozważania do jednego chromosomu, zakładając, że jakakolwiek segmentacja odnosi się do tego samego chromosomu. Wielkie litery służą do rozróżnienia segmentacji natomiast indeksy odnoszą się do kolejnych segmentów wzdłuż chromosomu czyli TAD-ów: $X[i], Y[j]$ (Rysunek 5a).

Definicja 2 *Na segmentach są zdefiniowane następujące funkcje:*

(i) początek segmentu: $s((a,b]) = a$,

(ii) koniec segmentu: $e((a,b]) = b$,

(iii) długość segmentu: $|(a,b]| = b - a$,

Jak wspomniano powyżej, wszystkie segmentacje $(X, Y, ...)$ odnoszą się do jednego chromosomu zatem $|X| = |Y| = ... = N$ oznacza jego długość.

Definicja 3 *Część wspólna dwóch segmentacji X i Y wyznacza segmentację $o_{X,Y}$ o następujących właściwościach:*

(i) $\forall_i \forall_j X[i] \cap Y[j] \neq \varnothing \implies X[i] \cap Y[j] \in o_{X,Y}$

(ii) $\forall_i \forall_{j>i} s(o_{X,Y}[i]) < s(o_{X,Y}[j])$

Dla uproszczenia przy rozważaniu 2 segmentacji X, Y indeks dolny w $o_{X,Y}$ można pominąć.

Definicja 4 *Funkcja $f_{X,Y}(i)$ zwraca oryginalny segment z segmentacji X, który zawiera segment $o[i]$. Bardziej formalnie:*

$$f_{X,Y}(i) = (a,b] \ \textit{ takie, że } \ (a,b] \in X \wedge o_{X,Y}[i] \subseteq (a,b]$$

Ta sama funkcja oznaczona przez $f_{Y,X}$ służy do wyszukiwania segmentów w drugiej segmentacji.

Definicja 5 (miara BP) *Niech X, Y będą 2 segmentacjami takimi, że $|X| = |Y|$. Wtedy miara BP jest określona w następujący sposób:*

$$d(X,Y) = 1 - \frac{1}{N} \sum_{i}^{n} \frac{|o[i]|^2}{\max(|f_{X,Y}(i)|, |f_{Y,X}(i)|)} \tag{1}$$

W pierwszej części tego rozdziału dowodzimy następującego twierdzenia:

Twierdzenie 1 *Funkcja $d(X,Y)$ jest metryką.*

W drugiej części skupiamy się na badaniu własności miary BP przy użyciu symulowanych i rzeczywistych danych Hi-C. W celu symulacji segmentacji A wylosowaliśmy ustaloną liczbę TAD-ów z rozkładu długości domen uzsykanego poprzez dopasowanie do danych rzeczywistych. Następnie, wygenerowaliśmy 3 rodzaje segmentacji B poprzez losowe przesunięcia granic domen segmentacji A według 3 scenariuszy, które imitowały różne schematy dopasowań (od podobnych do zupełnie różnych). Otrzymane pary segmentacji A, B poddaliśmy analizie podobieństwa przy pomocy miary BP oraz 2 innych metryk: odległości JI (z ang. Jaccard Index) i odległości VI (z ang. Variation of Information). Analiza podobieństwa par segmentacji wykazała ograniczenia miary JI w porównaniu do miar BP i VI. Wyniki uzyskane przy użyciu dwóch ostatnich miar są zbliżone, jednak uważamy, że definicja oraz wartości miary BP mają bardziej przejrzystą interpretację i lepiej dyskryminują pomiędzy rozważanymi schematami dopasowań segmentacji niż odległość VI, chociaż należy zaznaczyć, że ostatni efekt jest subtelny.

W kolejnej analizie wykorzystaliśmy publicznie dostępne dane Hi-C z 6 różnych linii komórkowych złożonych z 14 zestawów segmentacji TAD, wliczając replikaty poszczególnych eksperymentów. Połączyliśmy segmentacje w pary i przyporządkowaliśmy każdą parę do jednej z 2 kategorii: między-komórkowe lub jedna komórka. Badanie rozkładów odległości pomiędzy parami segmentacji pokazało, że wszystkie 3 metryki rozróżniają analizowane kategorie, to znaczy odległości w porównaniach segmentacji pomiędzy replikatami tej samej linii komórkowej są niższe niż pomiędzy różnymi komórkami. Jednak zarówno miara BP, jak i VI wykazują wyższą istotność statystyczną pomiędzy rozkładami odległości badanych kategorii niż JI. Co ciekawe, nasza analiza wskazuje na pary komórek, które wykazują znaczne niedopasowanie segmentacji TAD. Kontrastuje to z ogólnym poglądem na wysokie podobieństwo segmentacji domenowej pomiędzy liniami komórkowymi raportowane w licznych badaniach [Dixon et al., 2012, Dixon et al., 2015].

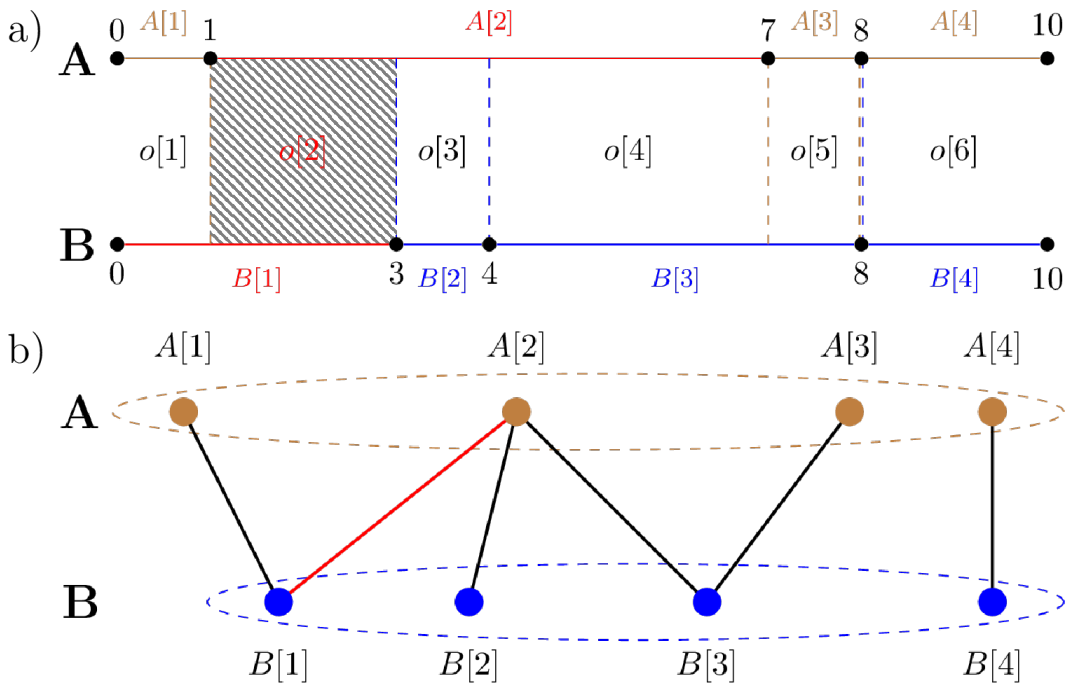
2.2 Wykrywanie różnicowych kontaktów Hi-C

Ostatni rozdział dotyczy problemu analizy różnicowej Hi-C. Nieformalnie celem takich badań jest wykrywanie oddziaływań chromatynowych dalekiego zasięgu, które są wzbogacone w danej linii komórkowej lub warunkach eksperymentalnych względem linii kontrolnej. Typowym problemem przy tego rodzaju analizach jest istnienie obciążeń spowodowanych różnicami w pokryciu odczytami oraz efektu zaniku liczby kontaktów wraz ze wzrostem odległości genomowej. Obydwa obciążenia maskują lokalne różnice i uniemożliwiają zastosowanie bezpośredniego porównania macierzy kontaktów. Typowe podejście do rozwiązania tego problemu polega na badaniu i modelowaniu zmienności interakcji za pomocą metod statystycznych.

Podczas naszych badań nad wykrywaniem różnic Hi-C analizowaliśmy profile liczby kontatków oraz ich zależności pomiedzy zbiorami danych Hi-C wykorzystując koncepcję sparowanych zbiorów interakcji $\mathcal{D}_k$ (z ang. paired interaction sets):

$$\mathcal{D}_k = \{(a_{ij}, b_{ij}) \mid k = |i - j|, a_{ij} \neq 0, b_{ij} \neq 0\}$$

gdzie: a_{ij}, b_{ij} są komórkami macierzy kontaktów A, B odpowiednio, a k odpowiada k-tej przekątnej. Przeprowadzone przez nas analizy ujawniły istotne korelacje profili interakcji pomiędzy parami eksperymentów przy ustalonych odległościach genomowych (Rysunek 6). Co zaskakujące, opisany trend utrzymuje się nawet dla map kontaktów różnych linii komórkowych. Chociaż tylko niewielka część przekątnych wykazuje znaczące korelację pomiędzy macierzami Hi-C, należy podkreślić, że większość kontaktów znajduje się w tym właśnie zakresie odległości genomowych. Oprócz wspomnianego dotychczas trendu zauważyliśmy



Rysunek 5. Segmentacje chromosomów. a) Dwie segmentacje i ich części wspólne. Przykładowe przecięcie dwóch domen oznaczonych kolorem czerwonym jest wskazane za pomocą kreskowania. b) Segmentacje można zilustrować za pomocą grafu dwudzielnego, którego węzły odpowiadają domenom, a krawędzie wskazują na istnienie części wspólnej danej pary domen. 2 zbiory wierzchołków grafu odpowiadają porównywanym segmentacjom A, B zaznaczonych przerywanymi liniami brązową (zestaw domen A) i niebieską (zestaw domen B). Czerwona krawędź reprezentuje część wspólną 2 domen wyróżnionych w punkcie a) za pomocą czerwonego koloru.

istnienie sparowanych interakcji odbiegających od ogólnego wzorca opisanego powyżej (Rysunek 6, zaznaczone czerwonym kolorem). Obserwacje te traktujemy jako potencjalne kontakty różnicowe. Podsumowując, budujemy nasz model zerowy w oparciu o następujące założenia:

- większość interakcji między dowolną parą zbiorów danych Hi-C do ustalonej maksymalnej odległości genomowej jest nieróżnicowa i skorelowana,
- związek częstości kontaktów pomiędzy warunkami jest liniowy i często heteroskedastyczny,
- interakcje różnicowe (jeśli występują) wyróżniają się jako odstające od opisanego wzorca.

Aby możliwie dokładnie uwzględnić obserwacje opisane powyżej, stosujemy regresję ujemną dwumianową, która może skutecznie modelować zależności liniowe w sytuacji gdy założenie stałej wariancji nie jest spełnione. Dodatkowo, aby złagodzić wpływ obserwacji odstających na proces estymacji parametrów, stosujemy metody odpornościowe (z ang. robust) opracowane przez [Aeberhard et al., 2014]. Przed estymacją parametrów grupujemy sparowane zbiory interakcji za pomocą kryterium odległości statystycznej. Mówiąc ściślej, traktujemy zbiory

$\mathcal{D}_k, \mathcal{D}_l$ (takie, że $l > k$) jako próby losowe z rozkładów dwuwymiarowych F_X, F_Y i testujemy następującą hipotezę:

$$\mathbf{H_0}\colon F_X = F_Y \text{ vs } \mathbf{H_1}\colon F_X \neq F_Y$$

za pomocą testu energii (z ang. energy test) opracowanego w [Székely et al., 2004]. Dodatkowo przyjmujemy następujące uproszczenie. Zamiast testować hipotezę $\mathbf{H_0}$: $F_k = F_{k+1} = ... = F_l$ wykonujemy osobny test dla każdego kolejnego i, zaczynając od $k+1$ dopóki hipoteza $\mathbf{H_0}$: $F_k = F_i$ nie zostanie odrzucona. W ten sposób uzyskujemy pulę p złożoną z przekątnych k do $l = i - 1$. Poprawność zastosowanego przybliżenia uzasadniamy istnieniem silnego zaniku liczby kontaktów w funkcji odległości genomowej (z ang. power law). W konsekwencji większość przekątnych $i < j < k$ spełnia następującą implikację: jeśli $\mathcal{D}_j$ jest istotnie przesunięte w stosunku do $\mathcal{D}_i$, to również $\mathcal{D}_k$ będzie przesunięte w stosunku do $\mathcal{D}_i$. Zasadniczo naszą metodę można podsumować jako 3-etapową procedurę:

1. najpierw agregowane są sparowane zbiory interakcji na podstawie podobieństwa rozkładów,
2. następnie dla każdego zagregowanego, sparowanego zbioru interakcji estymowane są parametry modelu regresji ujemnej dwumianowej,
3. w ostatnim etapie określana jest istotność statystyczna dla komórek macierzy (tj. par regionów) względem zbudowanego modelu.

Aby obliczyć istotność statystyczną wzbogacenia b_{ij} w stosunku do a_{ij}, wyznaczamy p-wartość za pomocą następującego wzoru:

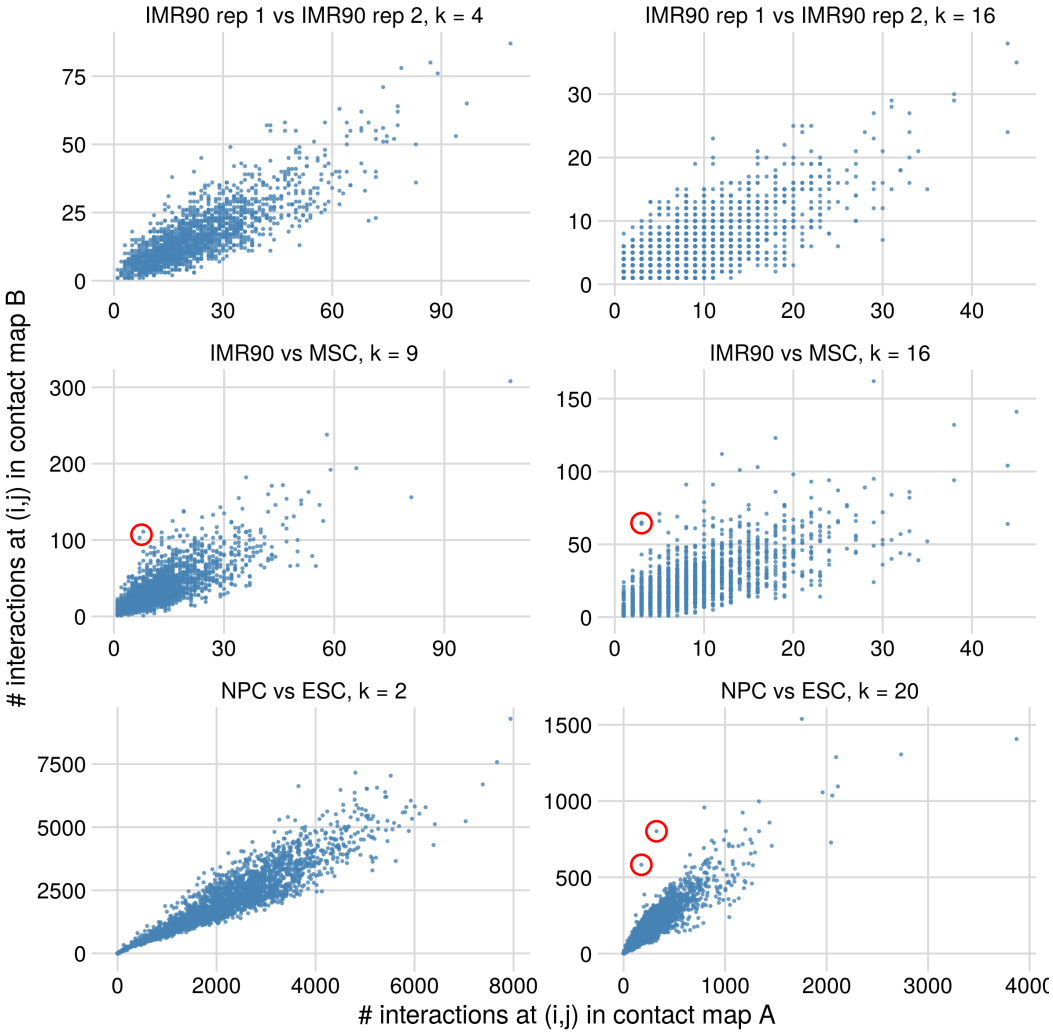
$$\mathrm{p}_{ij} = 1 - P(Y < b_{ij}), \text{ gdzie: } Y \sim NB(\hat{\beta}_p a_{ij}, \hat{\phi}_p)$$

w którym: $\hat{\beta}_p, \hat{\phi}_p$ to estymacje parametrów sparowanego zbioru interakcji $\widetilde{\mathcal{D}}_p = \mathcal{D}_k \cup \mathcal{D}_{k+1} \cup ... \cup \mathcal{D}_l$. Następnie, zbiory p-wartości są korygowane do q-wartości przy użyciu procedury Benajminiego-Hochberga. Ponadto po określeniu poziomu istotności wyniki mogą być przefiltrowane w celu zachowania najbardziej wiarygodnych interakcji różnicowych, które można pogrupować przy użyciu algorytmu wyszukiwania spójnych składowych grafu.

Na koniec postanowiliśmy sprawdzić skuteczność naszej metody, wykorzystując rzeczywiste i symulowane dane. Najpierw porównaliśmy p-wartości uzyskane z analizy różnicowej danych Hi-C różnych linii komórkowych z tymi pochodzącymi z replik bootstrap, a także z biologicznymi replikami z tych samych komórek. Zgodnie z oczekiwaniami zaobserwowaliśmy znacznie słabszy sygnał w dwóch ostatnich zbiorach w porównaniu do analiz przeprowadzonych na parach pochodzących z różnych linii komórkowych. Następnie postanowiliśmy sprawdzić jak skuteczna jest nasza metod (nazwana DiADeM) w porównaniu z dotychczas opracowanymi modelami do wykrywania kontaktów różnicowych [Lun and Smyth, 2015, Djekidel et al., 2018, Stansfield et al., 2018, Stansfield et al., 2019, Ardakany et al., 2019]. Do tego celu wykorzystaliśmy różne zbiory symulowanych danych Hi-C, w których lokalizacja kontaktów różnicowych jest znana. Wyniki przedstawione na rysunku 7 pokazują, że w przypadku większości typów danych symulowanych oraz intensywności interakcji różnicowych, DiADeM jest skuteczniejszy niż pozostałe metody, na co wskazują wyższe wartości PRAUC.

3 Podsumowanie

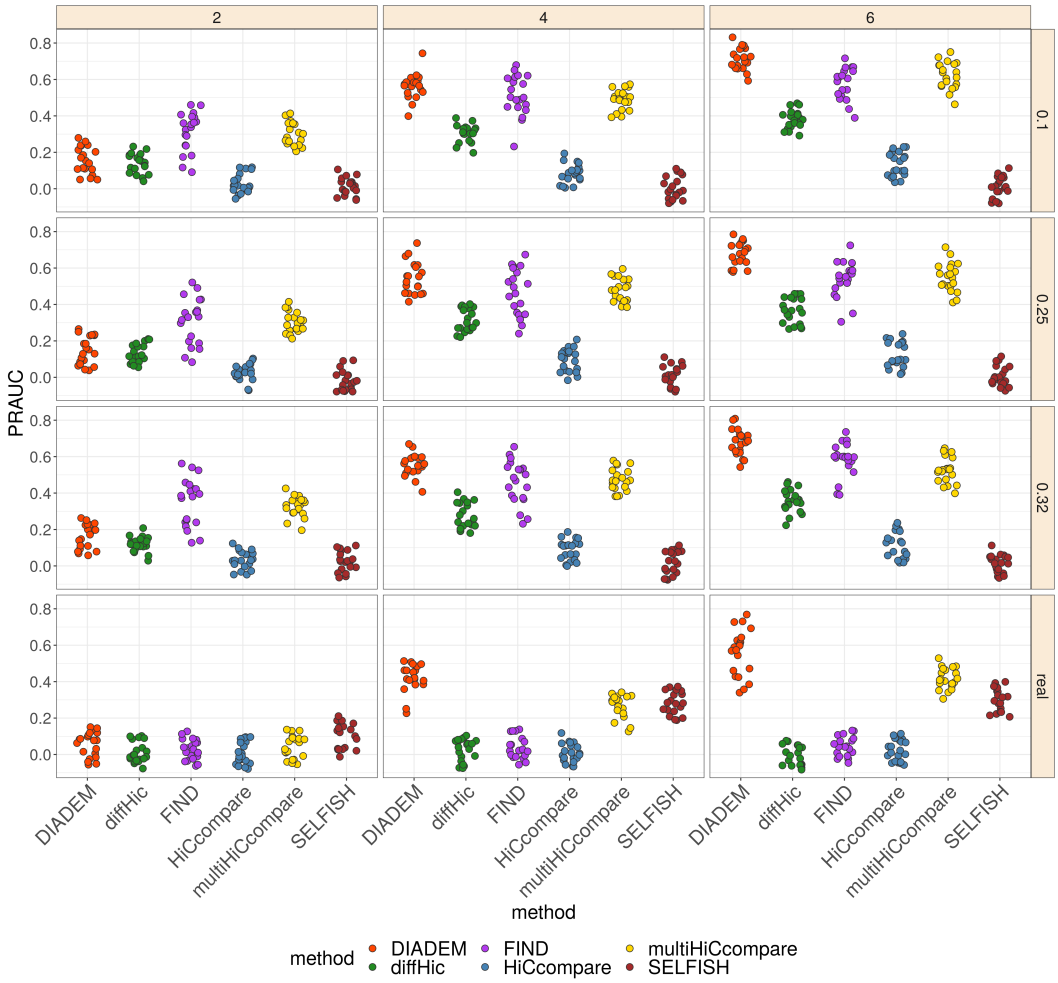
Niniejsza rozprawa jest poświęcona problemowi porównywania danych uzyskiwanych z eksperymentów Hi-C. W ramach rozprawy badano dwa podejścia do problemu, co zaowocowało



Rysunek 6. Zależności pomiędzy profilami liczby kontaktów różnych par macierzy Hi-C dla wybranych odległości genomowych k. Każdy punkt reprezentuje inną parę regionów. Oś pozioma wskazuje liczbę interakcji na mapie kontaktów A, podczas gdy oś pionowa wyraża liczbę interakcji w mapie kontaktów B. Obserwacje odstające zaznaczono czerwonymi okręgami.

stworzniem nowych metod odkrywania różnic w organizacji strukturalnej jąder komórkowych. Po pierwsze, zaproponowano nową miarę odległości dla porównań segmentacji chromosomów, zwaną odległością BP. Ta część pracy zawiera dowód właściwości metrycznych nowo wyprowadzonej funkcji. Analizy przprowadzone przy użyciu miary BP pokazują, że jest ona konkurencyjna w stosunku do podejść alternatywnych. Ponadto wykazano, że zastosowanie tej nowej metryki może prowadzić do ciekawych rezultatów odnośnie różnic w segmentacji chromatyny pomiędzy różnymi liniami komórkowymi.

Po drugie, badano problem odkrywania różnicowych oddziaływań chromatynowych.



Rysunek 7. Porównanie skuteczności różnych metod wykrywania różnicowych oddziaływań Hi-C mierzone polem powierzchni pod krzywą precyzja-czułość (z ang. precision-recall). Wiersze reprezentują dyspersję pomiędzy parami zestawów danych, a kolumny wskazują krotność zmiany pomiędzy wprowadzonymi interakcjami różnicowymi. Liczba symulowanych zbiorów danych wynosi 20.

Badanie zależności pomiędzy profilami interakcji różnych linii komórkowych ujawniło istotny poziom ich podobieństwa. Zjawisko to zastosowano do skonstruowania modelu statystycznego o nazwie DiADeM, który z powodzeniem wykrywa różnicowe kontakty chromatynowe. Analizy porównawcze pokazują, że model ten jest konkurencyjny w stosunku do istniejących metod, co sugeruje, że może posiadać potencjał w poszukiwaniu nowych interakcji regulatorowych.

Wykaz publikacji wyników przedstawionych w rozprawie:

Niskanen, Henri and Tuszynska, Irina* and **Zaborowski, Rafał*** and Heinäniemi, Merja and Ylä-Herttuala, Seppo and Wilczyński, Bartek and Kaikkonen, Minna U (2017). Endothelial cell differentiation is encompassed by changes in long range interactions between inactive chromatin regions. *Nucleic acids research*, 46(4):1724–1740. doi: 10.1093/nar/gkx1214.

Zaborowski, Rafał and Wilczyński, Bartek (2019). BPscore: an effective metric for meaningful comparisons of structural chromosome segmentations. *Journal of Computational Biology*, 26(4):305–314. doi: 10.1089/cmb.2018.0162.

Zaborowski, Rafał and Wilczyński, Bartek (under review). DiADeM: differential analysis via dependency modelling of chromatin interactions with robust generalized linear models. *bioRxiv*. doi: https://doi.org/10.1101/654699

Literatura

Aeberhard et al., 2014. Aeberhard, W. H., Cantoni, E., and Heritier, S. (2014). Robust inference in the negative binomial regression model with an application to falls data. *Biometrics*, 70(4):920–931.

Andrey et al., 2013. Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A switch between topological domains underlies hoxd genes collinearity in mouse limbs. *Science*, 340(6137):1234167.

Annunziato, 2008. Annunziato, A. (2008). Dna packaging: nucleosomes and chromatin. *Nature Education*, 1(1):26.

Ardakany et al., 2019. Ardakany, A. R., Ay, F., and Lonardi, S. (2019). Selfish: discovery of differential chromatin interactions via a self-similarity measure. *Bioinformatics*, 35(14):i145–i153.

Bai et al., 2013. Bai, J. P., Alekseyenko, A. V., Statnikov, A., Wang, I.-M., and Wong, P. H. (2013). Strategic applications of gene expression: from drug discovery/development to bedside. *The AAPS journal*, 15(2):427–437.

Cook et al., 2015. Cook, C. E., Bergman, M. T., Finn, R. D., Cochrane, G., Birney, E., and Apweiler, R. (2015). The european bioinformatics institute in 2016: data growth and integration. *Nucleic acids research*, 44(D1):D20–D26.

Cremer and Cremer, 2001. Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(4):292.

Cremer and Cremer, 2010. Cremer, T. and Cremer, M. (2010). Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3):a003889.

Dekker et al., 2002. Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science*, 295(5558):1306–1311.

Dixon et al., 2015. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331.

Dixon et al., 2012. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376.

Djekidel et al., 2018. Djekidel, M. N., Chen, Y., and Zhang, M. Q. (2018). Find: differential chromatin interactions detection using a spatial poisson process. *Genome research*, 28(3):412–422.

Flicek and Birney, 2009. Flicek, P. and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature methods*, 6(11s):S6.

Froussios et al., 2019. Froussios, K., Schurch, N. J., Mackinnon, K., Gierlinski, M., line Duc, C., Simpson, G. G., and Barton, G. J. (2019). How well do rna-seq differential gene expression tools perform in a complex eukaryote? a case study in arabidopsis thaliana. *Bioinformatics*, 1:6.

Gall and Pardue, 1969. Gall, J. G. and Pardue, M. L. (1969). Formation and detection of rna-dna hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences*, 63(2):378–383.

Lee et al., 2017. Lee, B. K. B., Tiong, K. H., Chang, J. K., Liew, C. S., Rahman, Z. A. A., Tan, A. C., Khang, T. F., and Cheong, S. C. (2017). Design: connecting gene expression with therapeutics for drug repurposing and development. *BMC genomics*, 18(1):934.

Lieberman-Aiden et al., 2009. Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293.

Lohmann et al., 2013. Lohmann, S., Herold, A., Bergauer, T., Belousov, A., Betzl, G., Demario, M., Dietrich, M., Luistro, L., Poignée-Heger, M., Schostack, K., et al. (2013). Gene expression analysis in biomarker research and early drug development using function tested reverse transcription quantitative real-time pcr assays. *Methods*, 59(1):10–19.

Lun and Smyth, 2015. Lun, A. T. and Smyth, G. K. (2015). diffhic: a bioconductor package to detect differential genomic interactions in hi-c data. *BMC bioinformatics*, 16(1):258.

Maston et al., 2006. Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59.

Rao et al., 2014. Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., et al. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.

Rudan et al., 2015. Rudan, M. V., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., and Hadjur, S. (2015). Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture. *Cell reports*, 10(8):1297–1309.

Sexton et al., 2012. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–472.

Stack et al., 1977. Stack, S. M., Brown, D. B., and Dewey, W. (1977). Visualization of interphase chromosomes. *Journal of cell science*, 26(1):281–299.

Stansfield et al., 2019. Stansfield, J. C., Cresswell, K. G., and Dozmorov, M. G. (2019). multihiccompare: joint normalization and comparative analysis of complex hi-c experiments. *Bioinformatics*.

Stansfield et al., 2018. Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I., and Dozmorov, M. G. (2018). Hiccompare: an r-package for joint normalization and comparison of hi-c datasets. *BMC bioinformatics*, 19(1):279.

Székely et al., 2004. Székely, G. J., Rizzo, M. L., et al. (2004). Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272.

Zheng and Xie, 2019. Zheng, H. and Xie, W. (2019). The role of 3d genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology*, page 1.

Zorn et al., 1979. Zorn, C., Cremer, C., Cremer, T., and Zimmer, J. (1979). Unscheduled dna synthesis after partial uv irradiation of the cell nucleus: distribution in interphase and metaphase. *Experimental cell research*, 124(1):111–119.

Zufferey et al., 2018. Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. *Genome biology*, 19(1):217.