

**RADA NAUKOWA DYSCYPLINY
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ**

zaprasza na
OBRONĘ ROZPRAWY DOKTORSKIEJ

mgr. Michała Własnowolskiego

która odbędzie się w dniu **30.10.2023 roku**, o godzinie **12:00** w trybie hybrydowym

Temat rozprawy:

„Computational Modelling and Analysis of the Three-Dimensional Structure of Human Genome at the Population Scale”

Promotor: prof. dr hab. Dariusz Plewczyński – Politechnika Warszawska

Recenzenci: dr hab. inż. Aleksandra Gruca, prof. uczelni – Politechnika Śląska

 prof. dr hab. Paweł Mackiewicz – Uniwersytet Wrocławski

 prof. dr hab. inż. Marta Szachniuk – Politechnika Poznańska

Obrona odbędzie się w Sali nr 40 w Gmachu Wydziału Matematyki i Nauk Informacyjnych Politechniki Warszawskiej.

Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Osoby zainteresowane uczestnictwem w obronie w formie zdalnej proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji dr. hab. Maria Ganzha ,
email : maria.ganzha@pw.edu.pl do dnia 28.10.2023 r. do godz. 23:59.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: www.bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-po-30-kwietnia-2019-r/Rada-Naukowa-Dyscypliny-Informatyka-Techniczna-i-Telekomunikacja/mgr-Michal-Wlasnowolski

Przewodniczący Rady Naukowej Dyscypliny
Informatyka Techniczna i Telekomunikacja
Politechniki Warszawskiej
dr hab. inż. Jarosław Arabas, prof. uczelni

Streszczenie

Przetwarzanie informacji biologicznej wewnątrz jądra komórkowego *metazoa* jest niezwykle złożonym procesem. Integruje ono wiele poziomów jej przechowywania i regulowania, takich jak sekwencja DNA, znaczniki epigenetyczne, elementy cis-regulacyjne oraz trójwymiarową strukturę chromatyny. Rosnące zapotrzebowanie na zaawansowane narzędzia obliczeniowe, umożliwiające poznanie złożonej organizacji genomu, zrozumienie różnic między populacjami oraz między komórkami osób zdrowych i chorych, jest bardziej aktualne niż kiedykolwiek. Pomimo iż obecnie dysponujemy zaawansowanymi metodami eksperymentalnymi pozyskiwania informacji o przestrzennych kontaktach chromatynowych, takimi jak ChIA-PET i Hi-C, ich zastosowanie wciąż wiąże się z wysokimi kosztami i jest czasochłonne, co ogranicza ich wykorzystanie w badaniach w skali populacji ludzkiej. W związku z tym, aby zmniejszyć koszty i zwiększyć dostępność badań nad przestrzenną organizacją genomu, niezbędny jest rozwój odpowiednich metod obliczeniowych.

W odpowiedzi na te potrzeby niniejsza praca prezentuje zaawansowane narzędzia informatyczne umożliwiające modyfikację wzorów kontaktów chromatynowych wynikających ze zmian sekwencji DNA, co umożliwia generowanie i porównywanie różnych modeli 3D odzwierciedlających zróżnicowanie populacyjne wariantów strukturalnych. To innowacyjne narzędzie zostało włączone do serwisu internetowego 3D-GNOME w wersji 2.0, umożliwiając unikalne badania trójwymiarowych struktur chromatyny dla tysięcy genomów ludzkich.

Dodatkowo, w celu zwiększenia wydajności obliczeń, opracowano narzędzie *cudaMMC*, które powstało na bazie algorytmu modelowania 3D-GNOME. Jest to metoda oparta na metodzie symulowanego wyżarzania Monte Carlo, rozbudowana o możliwość masowego zrównoleglania obliczeń na kartach graficznych (GPU). Pozwoliło to na znacznie szybsze generowanie trójwymiarowych struktur chromatyny (do 25 razy szybciej), przy jednoczesnym zachowaniu wysokiej jakości modeli.

W pracy przedstawiono również metodę obliczeniową służącą do tworzenia zespołów modeli 3D, zarówno dla struktur referencyjnych, jak i zmodyfikowanych przez warianty strukturalne. Ta nowatorska technika została zaimplementowana w wersji 3.0 serwisu internetowego 3D-GNOME. Umożliwia ona mapowanie enhancerów oraz promotorów genów na modele 3D, a

także obliczanie zmian w rozkładach odległości między tymi elementami regulatorowymi i genami w strukturach referencyjnych i zmodyfikowanych przez warianty. W celu obsługi generowania zespołów statystycznych modeli 3D oraz przetwarzania dużych zestawów danych, w serwisie 3D-GNOME zaimplementowano metodę *cudaMMC*. Obliczenia wykonano na klastrze Eden^N, będącym wewnętrznym heterogenicznym wysoko-wydajnym klastrem obliczeniowym HPC wyposażonym w węzły Nvidia DGX A100 i zarządzanym przez oprogramowanie kolejkowe Slurm.

Dzięki tym innowacjom, niniejsza praca dostarcza kompleksową platformę komputerową do badania wpływu wariantów strukturalnych na przestrzenną organizację genomu. Opisane narzędzia stanowią unikatowe źródło wiedzy pozwalającej na zrozumienie wpływu przestrzennej organizacji chromatyny na ekspresję genów, a także na badanie mechanizmów regulacji transkrypcji i chorób.

Abstract

Processing biological information within a metazoan cell nucleus is highly complex, as it must integrate multiple information storage and regulation layers such as DNA sequence, epigenetic marks, cis-regulatory elements, and the 3D structure of chromatin. The demand for advanced computational tools to unravel the intricate organisation of the genome, understand population differences, and discern between healthy and diseased cells is continually growing. While we have advanced experimental methods to obtain chromatin contacts, like ChIA-PET and Hi-C, their application is still costly and time-consuming, limiting their use in population-scale studies. This necessitates the adoption of computational approaches to reduce costs and increase accessibility.

Addressing the need for a sophisticated computational tool to apply changes to the chromatin contact pattern due to modifications of the underlying DNA sequence, this thesis introduces a comprehensive solution that facilitates generating and comparing distinct 3D models underpinned by Structural Variants (SV) driven changes. This innovative tool was incorporated into the 3D-GNOME 2.0 web service, enabling a unique exploration of chromatin 3D structures.

Moreover, to enhance the efficiency of these calculations and the manipulation of large chromatin models, this thesis presents the *cudaMMC* method. This method employs GPU-accelerated computing and the Simulated Annealing Monte Carlo approach, allowing for faster generation of chromatin 3D structures while maintaining model quality.

Furthermore, the study unveils a computational method designed to create ensembles of models for both reference and SV-altered structures. This novel technique, encapsulated within the 3D-GNOME 3.0 web server update, empowers researchers to map enhancers and gene promoters onto the 3D models. As a result, it's possible to calculate changes in the distribution of distances between these genomic features in reference and SV-affected structures. To handle the generation of 3D model ensembles alongside new large datasets, we implemented *cudaMMC* and established calculations on Eden^N high performance computing (HPC) cluster, an in-house heterogeneous computing resources equipped with Nvidia DGX A100 nodes and managed by Slurm. Through these advancements, this PhD thesis provides a comprehensive computational platform for studying the influence of structural variants on the genome's spatial organisation. These tools serve as a unique resource for understanding the effect of chromatin spatial organisation on genetic expression and investigating transcriptional regulation and disease mechanisms.

Recenzja Rozprawy Doktorskiej

Tytuł: Computational Modelling and Analysis of the Three-Dimensional Structure of Human Genome at the Population Scale

Autor: mgr Michał Własnowolski

Promotor: prof. dr hab. Dariusz Plewczyński

Tematyka badawcza

Rozprawa doktorska Michała Własnowolskiego przedstawia badania przeprowadzone przez Autora w zakresie bioinformatyki oraz genomiki obliczeniowej. Skupia się na modelowaniu struktury przestrzennej genomu ludzkiego oraz analizie różnic, jakie występują w tej strukturze pomiędzy osobnikami w skali populacyjnej. Oba te problemy stanowią kluczowe zagadnienie współczesnej biologii molekularnej. Ich rozwiązanie jest możliwe dzięki zastosowaniu szeregu technik laboratoryjnych, modelowania problemów z wykorzystaniem narzędzi badań operacyjnych oraz zaawansowanych, wysokowydajnych metod obliczeniowych – m.in. symulacji komputerowych oraz analizy danych – operujących na bardzo dużych zbiorach danych.

W ramach badań przedstawionych w rozprawie, opracowano zestaw narzędzi obliczeniowych umożliwiających modelowanie i analizę zmian trójwymiarowej struktury chromatyny, która jest głównym składnikiem chromosomów. W badaniach wykorzystano dane o kontaktach chromatynowych uzyskane z eksperymentów ChIA-PET, informacje o wariantach strukturalnych takich jak duże delecje, insercje, inwersje i duplikacje, a także modelowanie trójwymiarowej struktury chromatyny za pomocą silnika symulacyjnego 3D-GNOME, opartego o algorytm symulowanego wyżarzania Monte Carlo. Platforma obliczeniowa 3D-GNOME jest jednym z najważniejszych wyników badawczych uzyskanych w pracy doktorskiej i stanowi istotny wkład Autora w rozwój genomiki obliczeniowej.

Układ rozprawy doktorskiej (w tym informacje o jej poszczególnych częściach składowych)

Rozprawa doktorska oparta jest na cyklu czterech artykułów naukowych, z czego trzy ukazały się już drukiem, a czwarty został zgłoszony do czasopisma i jest w fazie recenzji. Rozprawa rozpoczyna się od streszczenia, które wprowadza w tematykę pracy, przedstawia główne osiągnięcia Autora i wskazuje możliwe dalsze kierunki rozwoju badań. Następnie, we wstępie do rozprawy Autor prezentuje motywację stojącą za rozwojem narzędzi obliczeniowych dedykowanych do modelowania i analizy struktur 3D chromatyny na skalę populacyjną. Sekcja 1.1 zawiera opis działania algorytmu 3D-GNOME, wykorzystywanego do modelowania struktur 3D w toku badań opisanych w cyklu publikacji. W sekcji 1.2 Autor przedstawia biologiczny kontekst problemu badania regulacji ekspresji genetycznej wewnątrz jądra komórkowego oraz potrzebę syntetycznej analizy, która integruje informację o czynnikach genetycznych, epigenetycznych oraz elementach cis-regulatorowych, takich jak promotory i enhancery. Mapowanie tych czynników na modele 3D umożliwia bardziej informatywną wizualizację wybranego *locus* oraz analizę zmian dystansów pomiędzy czynnikami regulatorowymi. W kolejnych sekcjach – 1.3 oraz 1.4 – Doktorant przedstawia podstawowe cele badań opisanych w rozprawie oraz listę artykułów ujętych w cyklu.

Drugi rozdział rozprawy poświęcony jest głównym osiągnięciom opisanym w publikacjach z cyklu. W zwięzły sposób Autor przedstawia w nim swój wkład w prace podsumowane publikacjami [P1]-[P4]. Na szczególną uwagę zasługują wykonane z dużą dbałością ilustracje oraz pseudokody umieszczone w tym rozdziale. Rozdział jest podzielony na cztery podrozdziały, z których każdy dotyczy jednej publikacji z cyklu. Układ treści jest podobny w każdym podrozdziale i zawiera krótką motywację do podjęcia badań, opis wykorzystanych metod oraz uzyskanych wyników badawczych. Prace są przedstawione w kolejności chronologicznej, poczynawszy od najstarszej, która ukazała się w 2019 r.

W trzecim rozdziale zaprezentowane są dodatkowe osiągnięcia Autora. Obejmują one m.in. udział w grantach naukowo-badawczych, wizyty akademickie oraz współautorstwo publikacji nieuwzględnionych w cyklu. W rozdziale czwartym Autor przedstawił kolejne cele, m.in. badanie trójwymiarowej struktury chromatyny genomów archaicznych populacji ludzkich, takich jak Neandertalczycy i Denisowianie. Badania te są obecnie realizowane przez Autora we współpracy międzynarodowej z zespołem Dr. Guya Jacobsa z University of Cambridge.

Dodatkowo w pracy znajdziemy spis treści, spis ilustracji, bibliografię, kopie czterech publikacji stanowiących osiągnięcie naukowe opisane w jednotematycznym cyklu,

oświadczenia współautorów tych publikacji oraz cztery publikacje nieujęte w cyklu wraz ze wskazaniem wkładu Doktoranta.

Układ pracy mgr-a Michała Własnowolskiego jest prawidłowy, typowy dla powszechnie przyjętego schematu rozpraw doktorskich opartych na cyklu publikacji naukowych.

Zastosowane piśmiennictwo

Zastosowane piśmiennictwo jest ściśle związane z przedmiotem badań Autora. Bibliografia zawiera 104 pozycje literaturowe. Autor rozprawy stosuje tzw. vancouver system cytowań. Odnosi się do artykułów publikowanych w najwyższej rangi czasopismach z dziedziny biologii, genomiki oraz bioinformatyki, m.in. *Nature*, *Nature Protocols*, *Nature Methods*, *Nature Communications*, *Nucleic Acids Research*, *Bioinformatics*, *Genome Biology*, *Proceedings of the National Academy of Sciences*. Znakomita większość tych publikacji to stosunkowo nowe (ukazały się w ciągu ostatnich 15 lat), wysoko cytowane prace. Przedstawiają one wyniki badań eksperymentalnych oraz obliczeniowych nad genomem, architekturą chromatyny lub opisują techniki eksperymentalne. Dobór bibliografii nie budzi zastrzeżeń. Wskazuje na bardzo dobre rozeznanie Doktoranta w tematyce podjętej w rozprawie doktorskiej oraz jego znajomość aktualnego stanu wiedzy w tym obszarze badawczym.

W spisie literatury znajdują się nieliczne usterki o charakterze redakcyjnym, na przykład nazwy czasopism przeważnie podawane są w formie skróconej lecz zdarzają się zapisy w formie pełnej mimo iż istnieje ogólnie przyjęty skrót (*Proceedings of the National Academy of Sciences* – poz. 2; *Current Opinion in Genetics Development* – poz. 4), nazwa tego samego czasopisma bywa pisana w różny sposób (*Genome Biol.* – poz. 48, 55; *Genome biology* – poz. 5, 16), zdarzają się nazwy własne pisane z małych liter (np. *dna* – poz 1, *science* – poz 8).

Cel pracy oraz zastosowane metody badawcze

Jak podaje Autor rozprawy, głównym celem pracy doktorskiej było opracowanie i wdrożenie bioinformatycznych narzędzi obliczeniowych do generowania i analizy trzeciorzędowych modeli chromatyny oraz zbadanie potencjalnego wpływu struktury przestrzennej chromatyny na aktywność genetyczną komórek. Przedmiotem prowadzonych badań był genom człowieka, jednak metody stworzone w ramach pracy można z powodzeniem zastosować do analizy innych genomów, w których tworzą się pętle chromatynowe.

Podczas prac badawczych Autor opierał się przede wszystkim na metodach szeroko stosowanych we współczesnej bioinformatyce łącząc przetwarzanie i modelowanie danych

biologicznych, analizy statystyczne, analizy dużych zbiorów danych, algorytmikę, algorytmy probabilistyczne, programowanie aplikacji internetowych, programowanie z wykorzystaniem kart graficznych, symulacje komputerowe.

Uważam, iż zastosowane metody badawcze są odpowiednie do rozwiązywanego problemu badawczego i wskazują na dobrą znajomość przez Autora rozprawy nowoczesnych i efektywnych metod oraz technologii stosowanych w naukach o życiu oraz naukach obliczeniowych. Założony przez Doktoranta cel pracy został osiągnięty.

Wyniki badań oraz ich praktyczne zastosowanie

Wyniki będące podstawą rozprawy doktorskiej zostały przedstawione w czterech publikacjach wieloautorskich [P1]-[P4]. Mgr Michał Własnowolski jest pierwszym autorem trzech spośród nich – [P2], [P3] i [P4]. Trzy publikacje – [P1], [P2], [P4] – ukazały się w wysoko punktowanych czasopismach naukowych z listy JCR w latach 2019-2023, czwarta została zgłoszona do czasopisma *Bioinformatics* i jest w trakcie recenzji.

Publikacja [P1], zamieszczona w czasopiśmie *Genome Biology* (IF₂₀₂₃ 18,01; 200 pkt MNiSW; kwartył Q1), prezentuje narzędzie opracowane do przewidywania zmian kontaktów chromatynowych wprowadzanych na podstawie wariantów strukturalnych. Za pomocą tego narzędzia przeprowadzono kompleksową analizę trójwymiarowej struktury ludzkiego genomu na skalę populacyjną. [P1] jest najlepiej cytowaną pracą Doktoranta (22 cytowania według Web of Science).

Artykuł [P2], opublikowany w *Nucleic Acids Research* (IF₂₀₂₃ 19,16; 200 pkt MNiSW; kwartył Q1), przedstawia implementację w serwisie internetowym 3D-GNOME (<https://3dgnome.mini.pw.edu.pl/>) metody opisanej w [P1] w ramach aktualizacji do wersji 2.0. Aktualizacja ta wprowadza narzędzia do porównywania zmian w kontaktach chromatynowych pomiędzy referencyjnym genomem GM12878 a alterowanym na podstawie wariantów strukturalnych. Umożliwia również porównywanie modeli trzeciorzędowej struktury referencyjnej i zmodyfikowanej, generowanych za pomocą silnika modelarskiego 3D-GNOME. Serwis został zintegrowany z zestawem danych wariantów strukturalnych 2504 genomów należących do 26 różnych populacji ludzkich z projektu 1000 Genome Project. Umożliwia on także wprowadzanie przez użytkowników własnych wariantów strukturalnych w formacie VCF. Publikacja [P2] ma 12 cytowań wg Web of Science.

Publikacja [P3] opisuje narzędzie cudaMMC, które jest rozszerzeniem silnika modelarskiego 3D-GNOME. Narzędzie to zwiększa wydajność obliczeń przez ich masowe zrównoleglenie na kartach GPU. Istotnie skraca to czas potrzebny do modelowania struktur 3D chromatyny – nawet do 25 razy dla największych chromosomów dla danych z *long-range* ChIA-PET CTCF. Największe przyspieszenie widać przy generowaniu całych kolekcji (*ensemble*) modeli 3D w oparciu o dane o znacznie większym rozmiarze, wygenerowane z eksperymentu *in situ* ChIA-PET, przy towarzyszącej temu większej stabilności czasu obliczeń. Wyniki te opisano w pracy, która została zgłoszona do czasopisma *Bioinformatics* (IF₂₀₂₃ 6,931; 200 pkt MNiSW; kwartył Q1) i jest w trakcie recenzji.

Artykuł [P4] opublikowany w *Nucleic Acids Research* (IF₂₀₂₃ 19,16; 200 pkt MNiSW; kwartył Q1), przedstawia narzędzie służące do analizy zmian rozkładu dystansów pomiędzy *loci*, które zawierają sekwencje promotorowe i enhancerowe. Analizator został dodany do serwisu internetowego 3D-GNOME w ramach aktualizacji do wersji 3.0, co wymagało wcześniejszej implementacji w tym serwisie narzędzia cudaMMC, opisanego w [P3]. Do obliczeń wykorzystywany jest klaster Eden, będący wewnętrznym heterogenicznym wysokowydajnym klastrem obliczeniowym HPC, wyposażonym w węzły Nvidia DGX A100. Prace te wymagały rozbudowy architektury serwisu internetowego i zarządzania obliczeniami przez oprogramowanie do kolejkowania zadań, *Slurm*. Dodatkowo baza danych z wariantami strukturalnymi została zaktualizowana do 3202 genomów z *1000 Genome Project*.

Na szczególną uwagę zasługuje fakt, iż wyniki badawcze uzyskane przez Doktoranta przyczyniły się do opracowania narzędzi analitycznych, które udostępniono poprzez serwis internetowy 3D-GNOME. Dzięki temu użytkownicy z całego świata mogą stosować je w praktyce w swoich badaniach naukowych. Serwis umożliwia badanie struktury 3D chromatyny przy wykorzystaniu zarówno z danych dostępnych na portalu (takich jak kontakty chromatynowe mediowane przez CTCF i RNAPII, warianty strukturalne z the 1000 Genome Project) jak i z własnych danych, które można wprowadzić do aplikacji. Dzięki przyjaznemu interfejsowi, platforma pozwala na przeprowadzanie skomplikowanych analiz nawet osobom nieposiadającym umiejętności programowania. 3D-GNOME umożliwia analizę wpływu zmian struktury 3D chromatyny na dystans między enhancerami a genami, a dzięki wykorzystaniu zrównoleglenia obliczeń na kartach GPU oraz infrastruktury klastra Eden, proces analizy jest wydajny. Autor rozprawy planuje rozbudowę bazy danych o kontakty chromatynowe dla kolejnych linii komórkowych (tj. H1ESC, HFFC6 i WTC11) oraz o dodatkowe warianty

strukturalne (the Simons Diversity Project), co pozwoli na badanie różnic między populacjami ludzkimi. Dodatkowo, planowane jest uwzględnienie danych dotyczących wymarłych ludzkich populacji, takich jak Neandertalczyki i Denisowianie, co poszerzy możliwości badawcze nad historią gatunku ludzkiego.

Uważam, iż wyniki badawcze uzyskane przez Doktoranta w rozprawie doktorskiej zasługują na wysoką ocenę. W szczególności doceniam fakt, iż mgr Własnowolski łączy umiejętności analityczne z algorytmicznymi, co pozwoliło na wdrożenie jego wyników badawczych w postaci ogólnodostępnych narzędzi obliczeniowych i umożliwienie efektywnego przetwarzania dużych wolumenów danych mających ogromne znaczenie we współczesnym świecie. Dużym atutem rozprawy są artykuły opublikowane w wysoko punktowanych czasopismach naukowych – publikacje [P1], [P2] i [P4] ukazały się w czasopismach z I kwartyła, ich sumaryczny współczynnik wpływu wynosi 56,33 a sumaryczna punktacja ministerialna to 600 pkt. Zgodnie z informacjami podawanymi przez Web of Science (na dzień 4.08.2023), mgr Michał Własnowolski posiada H-indeks = 4, a jego wszystkie publikacje były cytowane 48 razy (nie wliczając cytowań własnych).

Nieprawidłowości i braki w ocenianej rozprawie doktorskiej

Praca jest napisana ładnym i zrozumiałym językiem oraz przygotowana z dużą dbałością o szczegóły i zachowaniem wysokiej estetyki. Zauważyłam nieliczne błędy i nieścisłości, które nie wpływają na klarowność przekazu i nie zmieniają mojej wysokiej oceny rozprawy. Mają one często charakter błędów redakcyjnych, miejscami brakuje przedimków lub przecinków. Przykładowe usterki:

- str. 1: "To address the challenges" – lepiej brzmiałoby "To address these challenges"
- str. 5: "a singletons heatmap" – powinno być "a singleton heatmap"
- str. 5: "On each level of simulation Monte Carlo" – brak przecinka przed "Monte Carlo"
- str. 5: "represented by the form" – powinno być "represented by the formula"
- str. 5: "simulation level the energy" – brak przecinka przed "the energy"
- str. 6: "Next term represent binding energy E_b , defined as:" – powinno być "Next term represents binding energy E_b and is defined as"
- str. 6: "In energy form, w_s , w_b , w_o and w_h are energy terms weights." – to wyjaśnienie powinno znaleźć się bezpośrednio pod wzorem (1.5) lub należałoby napisać, że odnosi się ono do tego wzoru, np. „In formula (1.5), w_s , w_b , w_o and w_h denote weights of energy terms."

- str. 7: "This necessitates specific regulation of gene expression, including at its initial stage – transcription" – nietypowy szyk wyrazów w zdaniu powoduje, że przekaz jest nieco niejasny
- str. 7: "These sequences operate by spatially interacting" – "These sequences operate by spatial interactions"

Uważam również, iż nie jest potrzebne zapowiadanie co Autor napisze w kolejnym rozdziale oraz umieszczanie w rozdziale wstępu, w którym Autor uprzedza, co znajdzie się w tym rozdziale. Takie powtórzenia według mnie osłabiają pracę i zmniejszają jej atrakcyjność.

W pracy (w rozdziale 3) zabrakło według mnie informacji o prezentacji wyników przez Doktoranta. Przypuszczam, że niejednokrotnie przedstawiał swoje wyniki badawcze na seminariach czy konferencjach naukowych. Udział w konferencjach jest istotną częścią pracy każdego naukowca, a wystąpienia konferencyjne to równie istotna forma przedstawiania swoich wyników badawczych jak ich publikowanie w artykułach naukowych.

Wnioski końcowe

Autor pracy wykazał się umiejętnością poprawnej i przekonującej prezentacji wyników przeprowadzonych badań oraz trafnością wnioskowania. Dowiódł, iż w wysokim stopniu poznał dotychczasowy stan wiedzy o podejmowanym w pracy badawczej temacie, przedstawiany w przedmiotowej literaturze światowej. Posiada ogólną wiedzę teoretyczną w obszarze Informatyki, Bioinformatyki i Genomiki Obliczeniowej, wykazuje się umiejętnością samodzielnego prowadzenia pracy naukowej i zastosowania wiedzy w praktyce.

Recenzowana praca zawiera oryginalne rozwiązanie problemu naukowego. Uzyskane przez autora wyniki badań zostały opublikowane w wiodących, wysoko punktowanych czasopismach z dziedziny. Pracę oceniam bardzo wysoko. Ze względu na istotność uzyskanych wyników, szerokie podejście do rozwiązywanego problemu oraz bardzo dobre publikacje będące podstawą pracy, składam wniosek o jej wyróżnienie.

Stwierdzam, że praca mgr-a Michała Własnowolskiego pt. „Computational Modelling and Analysis of the Three-Dimensional Structure of Human Genome at the Population Scale” spełnia wymagania stawiane rozprawom doktorskim określone w art. 13.1 Ustawy o stopniach naukowych i tytule naukowym z dnia 14.03.2003 oraz stanowi oryginalne rozwiązanie przez autora zagadnienia naukowego.

.....
prof. dr hab. inż. Marta Szachniuk

WARSAW UNIVERSITY OF TECHNOLOGY

ENGINEERING AND TECHNOLOGY

Information and Communications Technology

PhD Thesis

Computational Modelling and Analysis of the
Three-Dimensional Structure of Human Genome
at the population scale

Michał Własnowolski, MA, BSc

Supervisor:

Prof. Dariusz Plewczyński, PhD

Warsaw 2023

Abstract

Processing biological information within a metazoan cell nucleus is highly complex, as it must integrate multiple information storage and regulation layers such as DNA sequence, epigenetic marks, cis-regulatory elements, and the 3D structure of chromatin. The demand for advanced computational tools to unravel the intricate organisation of the genome, understand population differences, and discern between healthy and diseased cells is continually growing. While we have advanced experimental methods to obtain chromatin contacts, like ChIA-PET and Hi-C, their application is still costly and time-consuming, limiting their use in population-scale studies. This necessitates the adoption of computational approaches to reduce costs and increase accessibility.

Addressing the need for a sophisticated computational tool to apply changes to the chromatin contact pattern due to modifications of the underlying DNA sequence, this thesis introduces a comprehensive solution that facilitates generating and comparing distinct 3D models underpinned by Structural Variants (SV) driven changes. This innovative tool was incorporated into the 3D-GNOME 2.0 web service, enabling a unique exploration of chromatin 3D structures.

Moreover, to enhance the efficiency of these calculations and the manipulation of large chromatin models, this thesis presents the *cudaMMC* method. This method employs GPU-accelerated computing and the Simulated Annealing Monte Carlo approach, allowing for faster generation of chromatin 3D structures while maintaining model quality.

Furthermore, the study unveils a computational method designed to create ensembles of models for both reference and SV-altered structures. This novel technique, encapsulated within the 3D-GNOME 3.0 web server update, empowers researchers to map enhancers and gene promoters onto the 3D models. As a result, it's possible to calculate changes in the distribution of distances between these genomic features in reference and SV-affected structures. To handle the generation of 3D model ensembles alongside new large datasets, we implemented *cudaMMC* and established calculations on Eden^N high performance computing (HPC) cluster, an in-house heterogeneous computing resources equipped with Nvidia DGX A100 nodes and managed by Slurm. Through these advancements, this PhD thesis provides a comprehensive computational platform for studying the influence of structural variants on the genome's spatial organisation. These tools serve as a unique resource for understanding the effect of chromatin spatial organisation on genetic expression and investigating transcriptional regulation and disease mechanisms.

Streszczenie

Przetwarzanie informacji biologicznej wewnątrz jądra komórkowego *metazoa* jest niezwykle złożonym procesem. Integruje ono wiele poziomów jej przechowywania i regulowania, takich jak sekwencja DNA, znaczniki epigenetyczne, elementy cis-regulacyjne oraz trójwymiarową strukturę chromatyny. Rosnące zapotrzebowanie na zaawansowane narzędzia obliczeniowe, umożliwiające poznanie złożonej organizacji genomu, zrozumienie różnic między populacjami oraz między komórkami osób zdrowych i chorych, jest bardziej aktualne niż kiedykolwiek. Pomimo iż obecnie dysponujemy zaawansowanymi metodami eksperymentalnymi pozyskiwania informacji o przestrzennych kontaktach chromatynowych, takimi jak ChIA-PET i Hi-C, ich zastosowanie wciąż wiąże się z wysokimi kosztami i jest czasochłonne, co ogranicza ich wykorzystanie w badaniach w skali populacji ludzkiej. W związku z tym, aby zmniejszyć koszty i zwiększyć dostępność badań nad przestrzenną organizacją genomu, niezbędny jest rozwój odpowiednich metod obliczeniowych.

W odpowiedzi na te potrzeby niniejsza praca prezentuje zaawansowane narzędzia informatyczne umożliwiające modyfikację wzorów kontaktów chromatynowych wynikających ze zmian sekwencji DNA, co umożliwia generowanie i porównywanie różnych modeli 3D odzwierciedlających zróżnicowanie populacyjne wariantów strukturalnych. To innowacyjne narzędzie zostało włączone do serwisu internetowego 3D-GNOME w wersji 2.0, umożliwiając unikalne badania trójwymiarowych struktur chromatyny dla tysięcy genomów ludzkich.

Dodatkowo, w celu zwiększenia wydajności obliczeń, opracowano narzędzie *cudaMMC*, które powstało na bazie algorytmu modelowania 3D-GNOME. Jest to metoda oparta na metodzie symulowanego wyżarzania Monte Carlo, rozbudowana o możliwość masowego równoległego obliczeń na kartach graficznych (GPU). Pozwoliło to na znacznie szybsze generowanie trójwymiarowych struktur chromatyny (do 25 razy szybciej), przy jednoczesnym zachowaniu wysokiej jakości modeli.

W pracy przedstawiono również metodę obliczeniową służącą do tworzenia zespołów modeli 3D, zarówno dla struktur referencyjnych, jak i zmodyfikowanych przez warianty strukturalne. Ta nowatorska technika została zaimplementowana w wersji 3.0 serwisu internetowego 3D-GNOME. Umożliwia ona mapowanie enhancerów oraz promotorów genów na modele 3D, a

także obliczanie zmian w rozkładach odległości między tymi elementami regulatorowymi i genami w strukturach referencyjnych i zmodyfikowanych przez warianty. W celu obsługi generowania zespołów statystycznych modeli 3D oraz przetwarzania dużych zestawów danych, w serwisie 3D-GNOME zaimplementowano metodę *cudaMMC*. Obliczenia wykonano na klastrze Eden^N, będącym wewnętrznym heterogenicznym wysoko-wydajnym klastrem obliczeniowym HPC wyposażonym w węzły Nvidia DGX A100 i zarządzanym przez oprogramowanie kolejkowe Slurm.

Dzięki tym innowacjom, niniejsza praca dostarcza kompleksową platformę komputerową do badania wpływu wariantów strukturalnych na przestrzenną organizację genomu. Opisane narzędzia stanowią unikatowe źródło wiedzy pozwalającej na zrozumienie wpływu przestrzennej organizacji chromatyny na ekspresję genów, a także na badanie mechanizmów regulacji transkrypcji i chorób.

Table of Contents

ABSTRACT	3
STRESZCZENIE	5
1. INTRODUCTION	11
1.1 3D-GNOME - MONTE CARLO SIMULATED ANNEALING METHOD FOR CHROMATIN 3D STRUCTURE MODELLING	13
1.2 BIOLOGICAL CONTEXT FOR GENE EXPRESSION REGULATION: THE RATIONALE BEHIND CHROMATIN 3D MODELLING	15
1.3 AIM AND RESEARCH THESES.....	17
1.4 PUBLICATIONS CONSTITUTING THIS DISSERTATION	19
1.5 LIST OF ADDITIONAL PUBLICATIONS NOT INCLUDED IN THE COLLECTION:.....	19
2. MAIN SCIENTIFIC CONTRIBUTION OF THE AUTHOR OF THE DISSERTATION	20
P1. SPATIAL CHROMATIN ARCHITECTURE ALTERATION BY STRUCTURAL VARIATIONS IN HUMAN GENOMES AT THE POPULATION SCALE	21
P2. 3D-GNOME 2.0: A THREE-DIMENSIONAL GENOME MODELING ENGINE FOR PREDICTING STRUCTURAL VARIATION-DRIVEN ALTERATIONS OF CHROMATIN SPATIAL STRUCTURE IN THE HUMAN GENOME	23
P3. <i>CUDAMMC</i> - GPU-ENHANCED MULTISCALE MONTE CARLO CHROMATIN 3D MODELLING	25
P4. 3D-GNOME 3.0: A THREE-DIMENSIONAL GENOME MODELLING ENGINE FOR ANALYSING CHANGES OF PROMOTER-ENHANCER CONTACTS IN THE HUMAN GENOME.....	27
SCIENTIFIC ACHIEVEMENTS	31
PARTICIPATION IN RESEARCH GRANTS	31
ACADEMIC VISITS.....	31
CONCLUSIONS AND FUTURE WORK.....	32
REFERENCES.....	34
COPIES OF THE PUBLICATIONS CONSTITUTING THE DISSERTATION	45
CO-AUTHORS' STATEMENTS	89
COPIES OF ADDITIONAL PUBLICATIONS NOT INCLUDED IN THE COLLECTION	108

This work has been supported by the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (TEAM to DP).

The „Three-dimensional Human Genome structure at the population scale: computational algorithm and experimental validation for lymphoblastoid cell lines of selected families from 1000 Genomes Project” project is carried out within the TEAM programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund.



Republic
of Poland



Foundation for
Polish Science

European Union
European Regional
Development Fund



1. Introduction

Chromatin, a complex of DNA and proteins, has a dynamic and highly intricate three-dimensional (3D) structure within the cell nucleus. This structure is fundamental in regulating gene expression and maintaining genomic integrity. However, characterising the 3D structure of chromatin is inherently more challenging than determining the structure of proteins. Unlike proteins, whose structures can be resolved through X-ray crystallography by crystallising and subsequently irradiating them with X-rays, chromatin cannot be studied using the same method due to its highly dynamic and flexible nature, which prevents it from being crystallised.

To address the challenges, scientists have been using experiments that probe chromatin structure in populations of cells or at the single-cell level. Techniques such as Hi-C and ChIA-PET capture the frequencies of physical contact between different nuclei chromatin regions. Hi-C, for instance, provides a broad view, capturing contacts mediated by various proteins. At the same time, ChIA-PET is more targeted and can identify contacts mediated by specific proteins, with particular emphasis on CTCF, RNAPII, and cohesin proteins due to their roles in building and modifying the spatial structure of chromatin. Single-cell techniques such as single-cell Hi-C and single-cell ChIA-PET have been developed to obtain a more detailed picture of chromatin interactions. However, they are currently limited by relatively low contact coverage.

A network of chromatin junctions is derived from these experimental data, represented by an arc diagram (Fig. 1). In this diagram, two chromatin regions in close proximity in 3D space are connected, and the height of the arc depends on the frequency of the contact's presence in cell nuclei.

Since these experimental techniques yield data on chromatin contacts rather than direct structural information, there is a need for computational models to translate these contact frequencies into meaningful 3D structures.

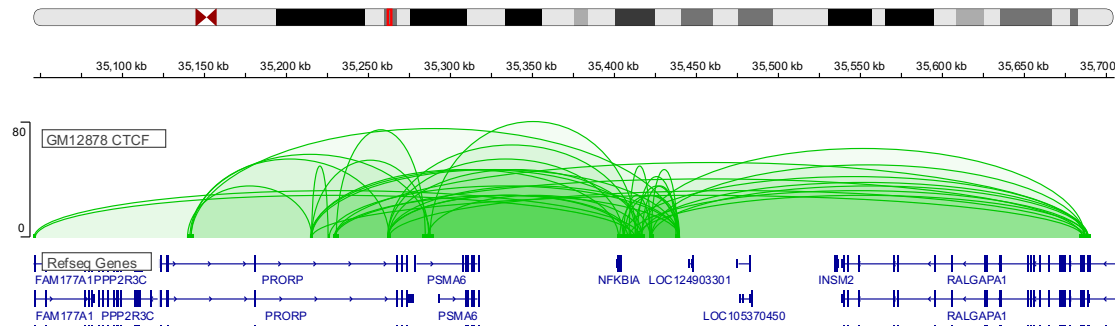


Figure 1 Arc diagram showing chromatin contacts mediated by CTCF protein in region chr14:35,045,500-35,704,867 for the GM12878 cell line

This is where the 3D-GNOME approach becomes relevant. 3D-GNOME is a computational method that utilises the contact data obtained from experiments like Hi-C and ChIA-PET to construct statistically plausible 3D models of chromatin structure. By doing so, it bridges the gap between contact information and spatial configuration, offering a valuable tool for investigating the three-dimensional organisation of chromatin and its implications for genomic functions. 3D-GNOME is one of several methods developed for this purpose and has been applied in the studies comprising this PhD thesis.

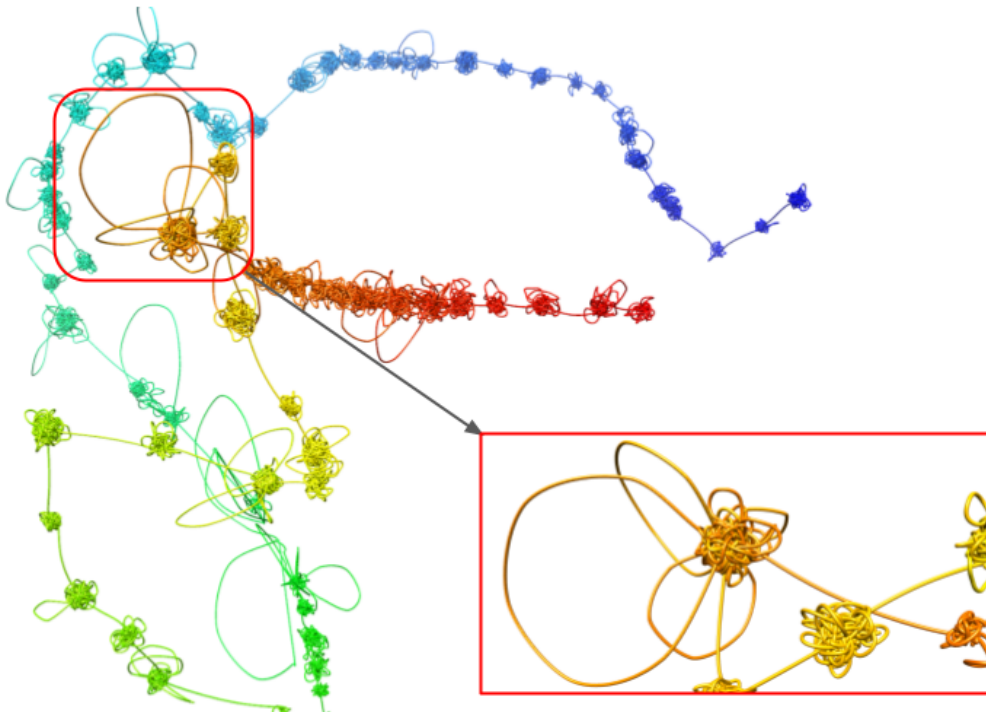


Figure 2 Example of a full-chromosomal 3D model of chromosome 1 from the GM12878 cell line, generated using the 3D-GNOME approach.

In the subsequent subsection, the 3D-GNOME approach will be described in detail, elucidating its effective use of chromatin contact data for reconstructing the three-dimensional architecture of chromatin.

1.1 3D-GNOME - Monte Carlo Simulated Annealing Method for Chromatin 3D Structure Modelling

3D-GNOME is a chromatin 3D structure modelling method based on a Monte Carlo simulated annealing algorithm. It uses a multiscale bead-on-a-string approach where each bead represents a particular region of chromatin on a different level of spatial organisation: nucleus, chromosome, domain, chromatin interaction anchor, and single chromatin loop (sub-anchor level). Relation between different levels is handled by a tree structure, with the nucleus as a root node and the following levels of genomic organisation in a parent-child relationship. Simulation proceeds from top to bottom: firstly, the algorithm performs low-resolution modelling to find chromosome positions in a nucleus sphere. After that, it identifies the positions of the topological domains of every single chromosome. The following steps perform high-resolution modelling to find positions of chromatin interaction anchors and sub-anchors, providing loops to the model structure. Beads with relation to different parent nodes are simulated independently, and the initial position of the child node (except sub-anchor simulation level) is found randomly in a sphere centred at the parent node.

3D-GNOME generates models based on paired-end chromatin interactions mediated by specific proteins like CTCF, RNAPII, or cohesin, derived from the ChIA-PET experiment, and considers two types of interactions. Low-resolution modelling uses singletons - chromatin contacts observed only once in a single experiment, and which heatmap shows a significant correlation with the Hi-C matrix. The inter-chromosomal singletons are used for chromosomal position modelling, whereas intra-chromosomal are used for domain modelling. The algorithm uses PET (paired-end tags) interaction clusters (high-frequent chromatin contacts) for high-resolution modelling, which consists of 2 stages. On the anchor level, each bead represents one of the anchors that create chromatin interaction. To create a chromatin loop, on the sub-anchor level, a preset number of beads is positioned between adjacent anchors based on both PET

clusters and singletons heatmap. PET clusters are used for splitting chromosomes into segments - of variable lengths, approximately 2Mbp (million base pairs) chromatin fragments that include topological domains extended by flanked regions. The differences between segment lengths are based on variances in the sizes of topological domains. The split might be performed automatically by clustering chromatin contacts or manually. In the last step, after modelling sub-anchor positions, the loop shape might be refined using a singletons heatmap that occurs within interaction and CTCF motif orientation (according to the loop extrusion model).

On each level of simulation Monte Carlo simulated annealing algorithm is used to find a structure that has minimum energy, represented by form:

$$E_{total}(\{\vec{r}_i\}) = \alpha E_{polymer}(\{\vec{r}_i\}) + \beta E_{data}(\{r_{ij}\}, \{d_{ij}\}),$$

where the first term denotes physical polymer constraints like stretching and bending, whereas the second term represents interactions gained from experimental data.

For chromosome, segment and anchor simulation level the energy function is

$$E = \sum_{i,j} (d_{ij} - r_{ij})^2,$$

where r_{ij} defines an actual distance between bead i and j , and d_{ij} defines preferred distance.

At chromosomal and segment levels, preferred distance is an inverse relationship

$$d_{ij} \sim c f_{ij}^{-\alpha},$$

with constant c and α , and f_{ij} as an interaction frequency between beads i and j . At the anchor level d_{ij} is defined as

$$d_{ij} = \delta + \alpha e^{-v(f_{ij}-\gamma)},$$

where δ , α , v and γ are constant. At sub-anchor level, energy consist of four terms:

$$E_{sub} = w_s E_s + w_b E_b + w_o E_o + w_h E_h.$$

In that equation, stretching term E_s is defined as

$$E_s = \sum_i \left(r_{i,i+1} - N_{i,i+1}^\beta \right)^2,$$

where $N_{j,j+l}$ is a genomic distance measure, and β is constant.

Next term represent binding energy E_b , defined as

$$E_b = \frac{1}{2} \sum_i (1 - \hat{v}_{i-1,i} \cdot \hat{v}_{i,i+1})^2,$$

where $\hat{v}_{i,i+1}$ is a unit vector pointing between sub-anchor beads i and $i+1$.

The third term includes CTCF motif orientation, according to the loop extrusion model,

$$E_o = \sum_{(i,j) \in P} (1 - \hat{o}_i \cdot \hat{o}_j),$$

where \hat{o}_i denotes on which chromatin strand CTCF motif is located, i and j are anchors that create chromatin interaction, and P is a set of interactions in one segment.

To simulate short-range singletons' impact on structure, the expected distances between each sub-anchor bead are calculated based on the physical constraints and modified by the singleton heatmap. These refined distances are used in the last energy term

$$E_h = \sum_{i,j} (d_{ij} - r_{ij})^2.$$

In energy form, w_s , w_b , w_o and w_h are energy terms weights.

1.2 Biological Context for Gene Expression Regulation: The Rationale Behind Chromatin 3D Modelling

Genetic information processing is a process of exceptional complexity, founded on chromatin - a heterogeneous polymer present in the nuclei of most types of eukaryotic cells.

Chromatin comprises two complementary DNA strands, composed of nucleotides adenine (A), thymine (T), cytosine (C), and guanine (G). The DNA strand is wound around histone proteins, forming specific and dynamically changing spatial structures.

Genes that encode the structure of proteins, synthesised in ribosomes based on transcribed copies of genes (mRNA), are encoded by the appropriate sequence of nucleotides. Various non-coding RNAs (ncRNAs), such as lncRNA, siRNA, miRNA, and eRNA, are also situated within the genome and play a significant role in regulating genetic expression.

In the case of metazoans, such as humans, the complete genetic information in each cell is identical (apart from mutations accumulated during development). However, the demand for products of genetic expression dynamically changes, not only due to the cell cycle and its interactions with the environment (e.g., interactions with pathogens), but also owing to the high degree of specialisation of cells into various tissues, such as muscle and nervous tissue. This gives rise to the need for specific regulation of genetic expression, including at the level of its initial stage - transcription.

The regulation of reading and processing genetic information is a multi-level phenomenon. The first level of regulation is represented by epigenetic markers such as cytosine methylation or histone modifications, such as acetylation or methylation. These markers are related to the level of chromatin condensation, as can be observed in the formation of active (euchromatin) and inactive (heterochromatin) chromatin compartments.

The second level of regulation involves the activity of proteins that bind to chromatin. These include, among others, transcription factors that recruit the RNAPII protein, which is responsible for transcription, i.e., the reading of genetic information. At this level, polycomb proteins, the CTCF protein, along with the cohesin ring, play a role in shaping the three-dimensional structure of chromatin. DHS proteins are also significant, influencing the availability of chromatin for transcription factors.

The third level of regulation encompasses cis-regulatory DNA sequences, including promoters, enhancers, and silencers. They operate by spatially interacting, for example, by bringing regions with an enhancer sequence closer to gene promoters, leading to the initiation of transcription of the RNAPII complex and increasing the level of transcription of the respective gene. For this interaction to occur, these sequences - sometimes situated several million base pairs apart on a DNA strand - must be in close proximity within three-dimensional space. Spatial chromatin organisation is crucial at this level of transcriptional regulation.



Figure 3 Ensemble of 3D models generated using the 3D-GNOME approach for the chromatin locus of the FXYP1 gene in the region chr17:63874585-64136737 for the MCF-7 breast cancer cell line. Among the 100 models, the one highlighted is the model closest to the average structure in the ensemble, with the gene promoter labelled in blue, the gene body in yellow, DNA methylation site cg23866403 (which is significant in breast cancer), and a potential enhancer region in orange.

Differences at each of these levels can be specific to individuals and populations and are also observed between healthy and diseased cells, such as cancer cells.

Understanding the mechanisms of gene expression and its regulation requires a comprehensive study of the interaction of individual genetic factors in space. This necessitates a synthetic analysis of differences at the levels of DNA sequences, epigenetics, and cis-regulatory elements, as well as an examination of their mutual interactions in three-dimensional space. Mapping individual elements onto a three-dimensional chromatin model is one of the methods enabling the analysis of interactions between them.

1.3 Aim and Research Theses

We are currently witnessing the dynamic development of technologies that enable the study of biological processes at the molecular level. The surge in genomic data has necessitated the parallel development of computational tools for their collection, processing, and analysis.

In this doctoral dissertation, my primary objective is to develop, implement, and co-create computational tools for generating and analysing three-dimensional chromatin models, as well as studying the potential impact of chromatin's spatial structure on the genetic activity of the cell. Although my work primarily centres on the human genome, the methods I have developed are versatile and can be applied to studying the genomes of other animals and those forming chromatin loops.

Below, I provide short descriptions of my work as detailed in the articles within this collection:

- In the first publication [P1], I introduced the impact of the three-dimensional structure of chromatin on genetic expression and presented a tool for predicting changes in the spatial structure based on structural variants (SVs).
- In the second publication [P2], I presented the integration of the tool described in the first part [P1] with the 3D-GNOME web service. This service offers a simple user interface and tools that enable the analysis of differences in the three-dimensional structure of chromatin, including chromatin contacts.
- In the third publication [P3], I introduced the *cudaMMC* tool, which is a new version of the 3D-GNOME method. This tool employs a modelling engine based on the Simulated Annealing Monte Carlo algorithm to generate three-dimensional chromatin structures. By utilising graphics processing units (GPUs) for computational acceleration, *cudaMMC* achieves up to a 30-fold increase in speed compared to the original version.
- In the fourth publication [P4], I presented an update to the 3D-GNOME web service, which includes integration with the *cudaMMC* tool. With this update, the process of generating 3D models has been optimised by offloading computations to the Eden^N computational cluster, which utilises Nvidia DGX A100 graphics cards, significantly accelerating calculations. Additionally, this update enables the generation of not just single models, but entire ensembles of 3D models. Furthermore, additional tools for analysing changes in distances between transcription enhancers and gene promoters in three-dimensional models have been included.

1.4 Publications constituting this Dissertation

- [P1] Sadowski, M., Kraft, A., Szalaj, P., **Wlasnowolski, M.**, Tang, Z., Ruan, Y., & Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome biology*, 20(1), 1-27.
IF=18.01, MNiSW points: 200
- [P2] **Wlasnowolski, M.**, Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y. & Plewczynski, D. (2020). 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Research*, [online] 48(W1), pp.W170–W176.
IF=19.16, MNiSW points: 200, related to ITT discipline
- [P3] **Wlasnowolski, M.**, Grabowski, P., Roszczyk, D., Kaczmarek, K., & Plewczynski, D. (2023). cudaMMC - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling. bioRxiv. <https://doi.org/10.1101/2023.06.12.544609>
in revision in *Bioinformatics* (IF=6.931, MNiSW points: 200, related to ITT discipline)
- [P4] **Wlasnowolski, M.**, Kadlof, M., Sengupta, K., & Plewczynski, D. (2023). 3D-GNOME 3.0: a three-dimensional genome modelling engine for analysing changes of promoter-enhancer contacts in the human genome. *Nucleic Acids Research*, gkad354.
IF=19.16, MNiSW points: 200, related to ITT discipline

1.5 List of additional publications not included in the collection

Herman-Izycka, J., **Wlasnowolski, M.**, & Wilczynski, B. (2017). Taking promoters out of enhancers in sequence based predictions of tissue-specific mammalian enhancers. *BMC medical genomics*, 10, 17-26.
IF=3.622, MNiSW points: 100

Sarkar, J. P., Saha, I., Rakshit, S., Pal, M., **Wlasnowolski, M.**, Sarkar, A., Maulik, U., & Plewczynski, D. (2019, July). A new evolutionary rough fuzzy integrated machine learning technique for microRNA selection using next-generation sequencing data of breast cancer. In Proceedings of the *Genetic and Evolutionary Computation Conference Companion* (pp. 1846-1854).

MNiSW points: 140, related to ITT discipline

Saha, I., Rakshit, S., **Wlasnowolski, M.**, & Plewczynski, D. (2019, October). Identification of epigenetic biomarkers with the use of gene expression and DNA methylation for breast cancer subtypes. In *Tencon 2019-2019 Ieee Region 10 Conference (Tencon)* (pp. 417-422). IEEE.

MNiSW points: 20, related to ITT discipline

Sarkar, J. P., Saha, I., Lancucki, A., Ghosh, N., **Wlasnowolski, M.**, Bokota, G., Dey, A., Lipinski, P., & Plewczynski, D. (2020). Identification of miRNA biomarkers for diverse cancer types using statistical learning methods at the whole-genome scale. *Frontiers in Genetics*, 11, 982.

IF=4.772, MNiSW points: 100

2. Main scientific contribution of the Author of the Dissertation

Next, I have showcased, through the example of publications collected in the compilation, the primary developmental work aimed at addressing critical scientific problems, such as 3D modelling at the population scale and the integration of chromatin spatial data with other regulatory genomic features for the investigation of gene expression regulation.

P1. Spatial chromatin architecture alteration by structural variations in human genomes at the population scale

Genome biology | MNiSW points: 200 | IF: 18.01

Background. Approximately 20 million base pairs, constituting 0.6% of the human genome, undergo structural variations (SVs) such as deletions, duplications, insertions, and inversions. Traditionally, the focus of studying SVs has been on their role in altering gene copy numbers and structures, which has implications for conditions such as cancer, intellectual disabilities, and various other health issues. Nonetheless, it is crucial to note that the majority of structure variations are situated within non-coding regions.

Certain SVs within these non-coding regions influence genomic sites recognised by proteins integral to organising the spatial structure of the genome within the cell nucleus. At present, catalogues of structural variants are being created. An example of this is the 1000 Genome Project consortium, which in release 3 identified over 2,500 structural variants in human genomes from 26 different human populations. However, we do not have such extensive data describing the spatial structure of chromatin, due to the high complexity and cost of the experiments used to generate them. That's why it is important to develop computational tools for predicting such chromatin structures ([1]–[67]).

Results. This study introduces a novel computational method for predicting alterations in chromatin contacts (Fig. 4), employing high-quality ChIA-PET data in tandem with population-scale SV data ascertained by the 1000 Genomes Consortium. This is the inaugural genome-wide analysis examining the influence of SVs on the three-dimensional structure of the human genome. Furthermore, the study models how SVs prompt changes in 3D genomic structures throughout the human population.

Significantly, the study brings to light the intricate relationship between genomic interactions and SVs and underscores the substantial impact of genetic variants on the higher-order organisation of chromatin folding. It also provides invaluable insights into the regulatory

mechanisms of gene transcription at a population level. Moreover, the study utilises the 3D-GNOME approach to construct 3D structures based on chromatin interactions altered by SVs.

One of the key focuses of this study is on chromatin interactions associated with enhancer regions and gene promoters, as these interactions play a pivotal role in regulating gene transcription. Through this focus, the study provides a deep understanding of how structural variations can affect the 3D structure of chromatin and emphasises the importance of chromatin architecture in genome regulation.

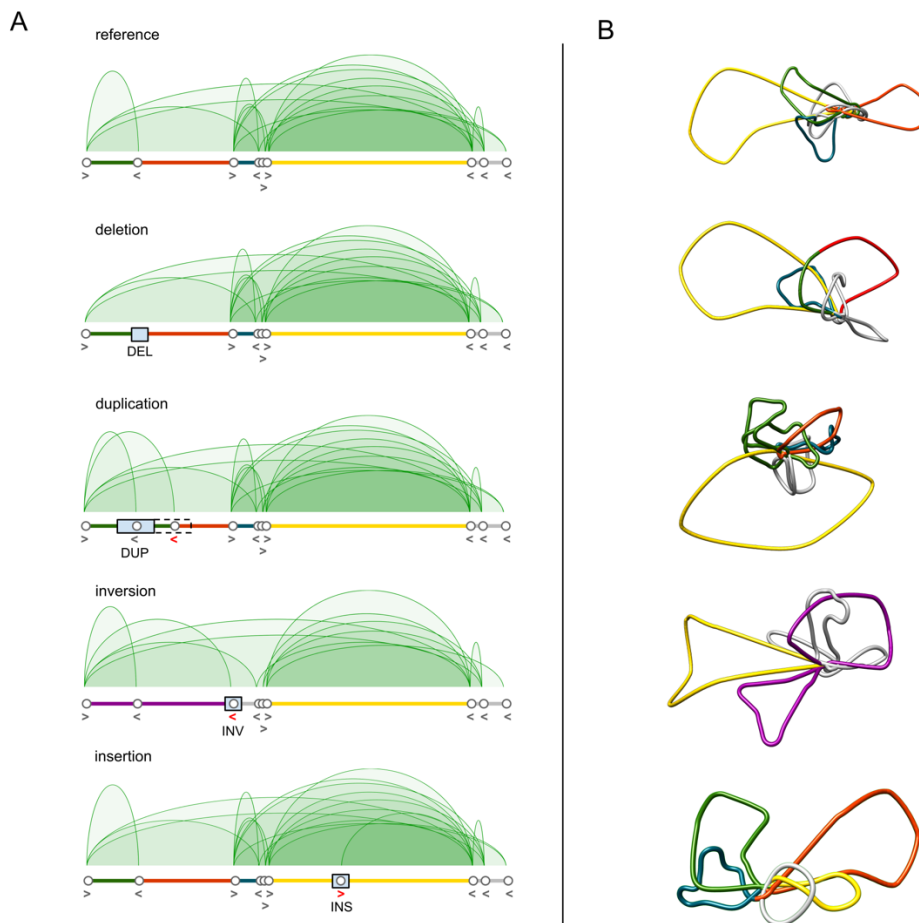


Figure 4 Diagrams illustrating the operation of the computational algorithm for predicting changes in 3D chromatin contacts caused by SVs. (A) CTCF ChIA-PET contact diagrams for a sample region chr1:47656996-48192898, which includes the TAL1 locus, are shown for the reference genome (GM12878) and upon the incorporation of SVs. The changes in patterns of CTCF-mediated contacts upon the addition of DUP, DEL, INV, or INS to the genomic sequence are displayed. SVs are indicated by blue rectangles. CTCF anchors and their directionality are represented by white circles and arrows, respectively. Alterations to the CTCF anchors induced by SVs are depicted by red arrows. (B) 3D models of CTCF-mediated chromatin structures corresponding to the genomic regions shown in (A). Loops are colour-coded to match the genomic regions represented below the CTCF contact diagrams in (A).

The algorithm for predicting changes in chromatin contacts was implemented in the 3D-GNOME web service update and described in [P2] article.

P2. 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome

Nucleic Acids Research | MNiSW points: 200 | IF: 19.16

Motivation. To further advance our research endeavours, I have chosen to provide a tool for predicting alterations in the spatial configuration of chromatin via the 3D-GNOME web service. In version 1.0, this service was used to generate three-dimensional models of chromatin using the Monte Carlo Simulated Annealing method, based on Chromatin Conformation Capture (3C) data. The service also provided tools for visualisation and analysis of 3D structures. Users could examine a specific DNA region by specifying the coordinates of the locus of interest. The service's architecture is built using the Flask development platform. Queries are saved to a MySQL database. The computations are written in Python, and external software is written in C++, PHP, and R. 3D-GNOME generates chromatin contact diagrams, plots, statistics, and 3D models of chromatin. Display and analysis of models were presented through an interactive viewer implemented in WebGL ([4], [5], [7], [8], [11], [13]–[17], [24], [68]–[75]).

Results. With the update to version 2.0 of 3D-GNOME (Fig. 5), described in this article, users have gained the ability to predict changes in the three-dimensional conformation of chromatin for a selected region of the human genome. Changes are introduced based on the tool described in [P1], using reference interaction data of CTCF and RNAPII from the ChIA-PET experiment for the GM12878 lymphoblastoid cell line, as well as structural variants found in the 1000 Genomes dataset.

Users can also input custom structural variants in VCF format. The web service has been upgraded to be compatible with Python 3.6+. The output has been expanded with a user-friendly interface, based on the Bootstrap 3 library, enabling easy manipulation and comparison of results from arc diagrams and statistics. The display of arc diagrams has been enhanced with the highlighting of regions where structural variants are applied.

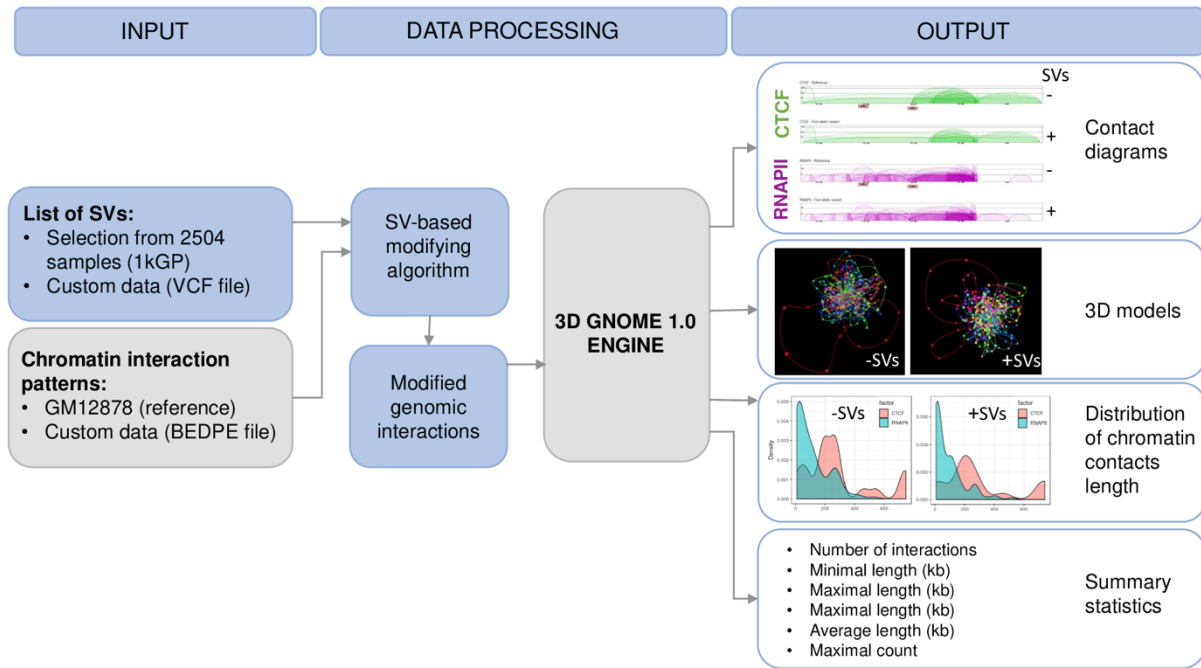


Figure 5 A diagram illustrating the workflow of 3D-GNOME 2.0.

Additionally, a download section has been created where users can download the results of the conducted analysis, including 3D models converted from 3D-GNOME's native .hcm format to PDB, which can be visualised using widely available 3D viewers like UCSF Chimera. To reduce calculation time, interaction data for 2,504 genomes in both allelic variants have been cached. Furthermore, the web service has been expanded to handle new datasets, such as structural variants from the 1000 Genomes Project and custom SVs in VCF format, integrated with existing data formats. It processes new data like VCF, applying it to the analysis pipeline, implementing structural changes, and displaying results in an easily accessible manner.

P3. *cudaMMC* - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling

bioRxiv (in revision in: Bioinformatics | MNiSW points: 0/200 | IF: 6.931)

Motivation. The next stage was to enhance the performance of generating 3D chromatin structures. With the recent advancements in 3C-based sequencing techniques, such as ChiA-PET and Hi-C, the volume of data generated has surged exponentially. This increase has made generating large 3D structures within a reasonable timeframe challenging, especially when creating model ensembles essential for conducting statistical analyses of 3D structure diversity of the human genome at the population scale. Such analyses are crucial for understanding chromatin's 3D structure's impact on transcriptional regulation and investigating changes in distances between regulatory elements, such as enhancers and promoters located on chromatin. This complexity underscores the importance of addressing the time performance of modelling with tools like 3D-GNOME. Consequently, we decided to expedite this process by distributing calculations across GPU cards, enabling a more efficient investigation into these complex chromatin structures ([69], [71], [76]–[84]).

Methods. The acceleration of the method is a result of a massive parallel search in the configuration space. The capacity of GPUs to execute thousands of independent threads concurrently offers an advantage over CPU processors for tasks that can be subdivided into numerous smaller threads. In the case of simulated annealing, it is challenging to achieve a balance between parallel and sequential computations because the global energy is calculated after local bead matching improvement.

Results. In our comparative evaluation, the *cudaMMC* algorithm demonstrated notable superiority over the 3D-GNOME method in terms of efficiency and speed in 3D chromatin structure modelling (Fig. 6). Operating on an NVIDIA Pascal GPU, the *cudaMMC* method achieved a remarkable speed-up of 3x to 25x in individual chromosome modelling, contingent upon chromosome size. Most striking was the substantial acceleration witnessed in the generation of model ensembles. When deploying large datasets from *in situ* ChIA-PET (a method that gains higher resolution, accuracy and volume data) to generate ensembles of 100

models, the *cudaMMC* algorithm successfully slashed computation times: chromosome 21 modelling was reduced from 85 minutes to about 11 minutes, chromosome 14 from approximately 8.5 hours to nearly 38 minutes, and chromosome 1 from around 3 days to just 2 hours. Moreover, by analysing the coefficient of variation for algorithm performance, *cudaMMC* shows much greater stability in comparison to 3D-GNOME.

Finally, I have integrated a tool that converts the output into the *mmCIF* format for broader usability, ideal for high-resolution whole chromosome models. In essence, our findings endorse the *cudaMMC* algorithm as a significant performance-enhancer in 3D chromatin structure modelling, underscoring its value to researchers in this field. The pseudocode illustrating the behaviour of *cudaMMC* is shown in Algorithm 1.

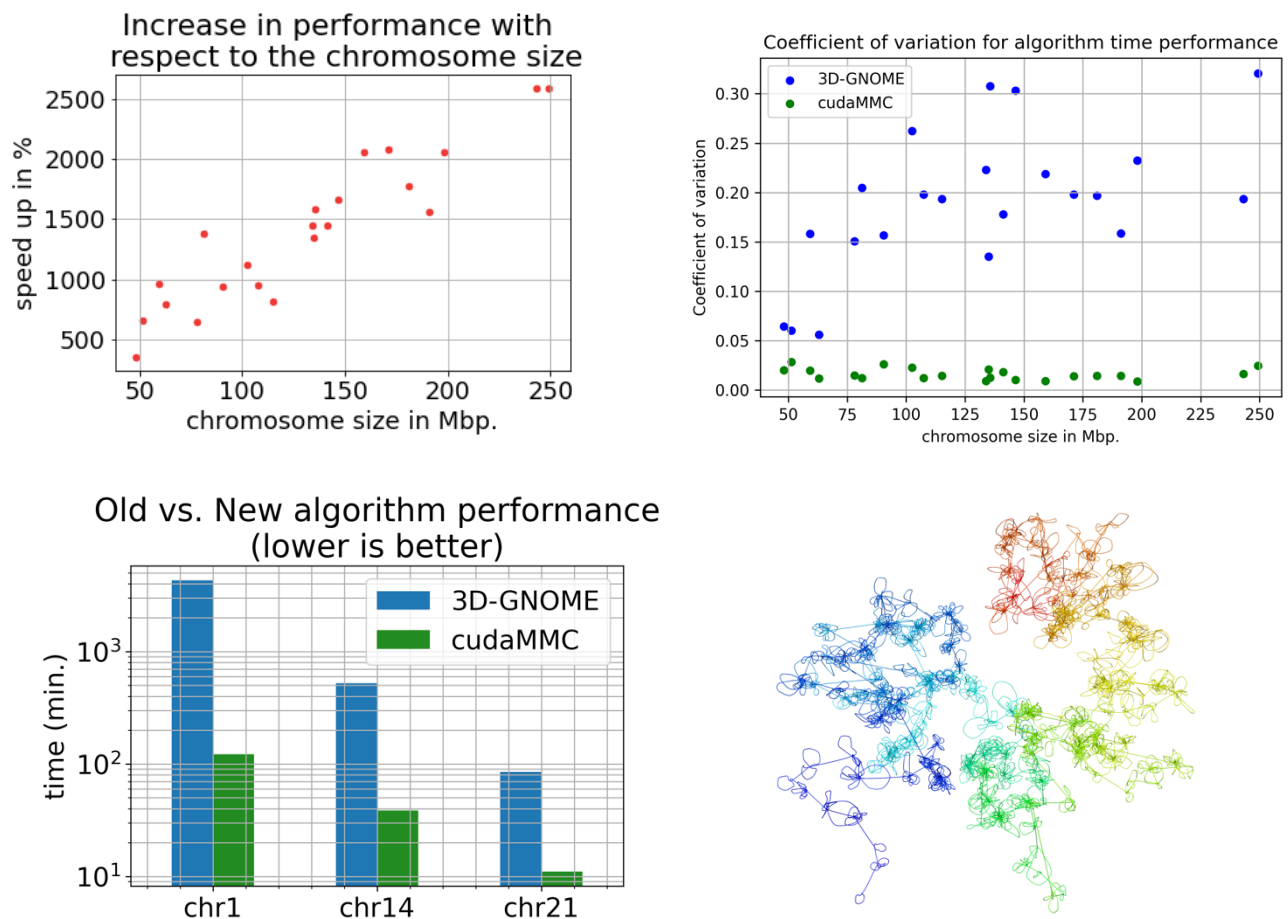


Figure 6 Comparison of 3D-GNOME and *cudaMMC* performance; full-chromosomal 3D model of chr1 generated using *cudaMMC*

Algorithm 1 Optimize Beads Locations

```
1: procedure OPTIMIZE_BEADS_LOCATIONS(volatile bool* optimum_found, double temp, half3  
   beads_positions[], T1 beads_property1[], T2 bead_property2[], ...)  
2:   for all bead in beads_positions, do in parallel warps (warpIdx) do  
3:     while not optimum_found do  
4:       bead  $\leftarrow$  beads_positions[warpIdx]  
5:       energy  $\leftarrow$  calculate_bead_energy(bead_positions, warpIdx)  
6:       for all thread do in a loop, i in range(512) do  
7:         x, y, z  $\leftarrow$  get_random_values()  
8:         moved_bead  $\leftarrow$  bead + [x, y, z]  
9:         new_energy  $\leftarrow$  calculate_energy_after_moving_bead(bead_positions, bead, warpIdx)  
10:        if new_energy < energy or evaluate_chance_prop_to_temp(temp, new_energy - energy) then  
11:          swap(bead, moved_bead)  
12:          swap(energy, new_energy)  
13:        end if  
14:      end for  
15:      decrease_temp(temp)  
16:      lowest  $\leftarrow$  warp_reduce_min_sync(energy)  
17:      if energy == lowest and lowest < calculate_bead_energy(warpIdx) then  
18:        beads_positions[warpIdx]  $\leftarrow$  bead  
19:        if optimum_found() then  
20:          optimum_found  $\leftarrow$  true  
21:        end if  
22:      end if  
23:    end while  
24:  end for  
25: end procedure
```

Algorithm 2 Calculate Energy After Moving Bead

```
1: procedure CALCULATE_ENERGY_AFTER_MOVING_BEAD(bead_positions, idx, warpIdx)  
2:    $\triangleright$  calculates energy substituting bead_position for beads_positions[idx]  
3:   substituted_positions  $\leftarrow$  bead_positions with bead_position at index idx  
4:   return calculate_energy(substituted_positions, warpIdx)  
5: end procedure
```

Description:

warpIdx => the same for all contiguous 32 CUDA.THREADS, 0-31

threadIdx => index in grid, unique for each thread, 0-NUM.THREADS

temp => temperature of simulated annealing

beads_positions => array of 3D structures containing bead positions

Algorithm 1 Pseudocode showing the cudaMMC approach behaviour

P4. 3D-GNOME 3.0: a three-dimensional genome modelling engine for analysing changes of promoter-enhancer contacts in the human genome

Nucleic Acids Research | MNiSW points: 200 | IF: 19.16

Motivation. To address the challenges faced in human genetics, particularly in deciphering genetic expression regulation and understanding the impact of genome variability on transcription, a potent computational approach is essential. With tools at our disposal for rearranging SV-driven chromatin 3D structures (described in [P1] and [P2]), and a GPU-accelerated 3D modelling engine (described in [P3]), our subsequent step was to devise a computational method for analysing changes in distance distributions within model ensembles, between reference and SV-altered structures. The aim was to integrate this methodology into the 3D-GNOME 3.0 web server update, equipping researchers with the ability to map enhancers and gene promoters onto the 3D models. The Scheme of 3D-GNOME 3.0 server architecture is presented on Figure XYZ ([5], [17], [68], [69], [74], [76], [85]–[103]).

Results. The first step I did was to integrate new, much larger, high-resolution datasets of in situ ChIA-PET chromatin interactions with higher confidence, as well as updated structural variants from the 1000 Genome Project datasets of 3,202 samples, by including 30x high-coverage data.

To handle these new datasets, we replaced the 3D modeller engine with *cudaMMC* and, to further capitalise on its GPU acceleration, we migrated the calculations of the 3D-GNOME web server to Eden's cluster, a high-performance computing (HPC) cluster. Eden's cluster is an in-house heterogeneous computing resource equipped with Nvidia DGX A100 nodes and is deployed at the Faculty of Mathematics and Information Science at Warsaw University of Technology. For security reasons, we segregated elements of the new architecture; the web server interface runs in an LXC container within a Proxmox environment, and the calculations

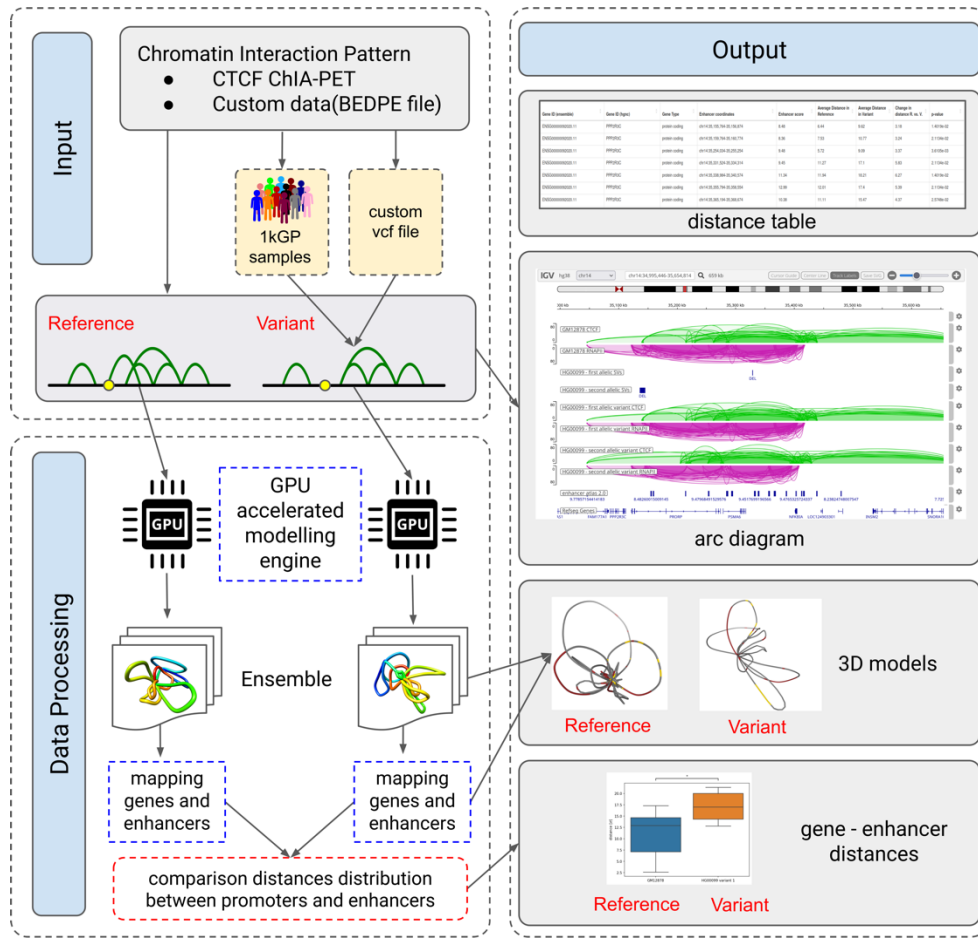


Figure 7 Diagram of 3D-GNOME 3.0 Server Architecture

are managed by Slurm. The schema of the 3D-GNOME architecture is described in Figure 7 and integration with the cluster is shown in Figure 8.

In this environment, the computational power was sufficient to establish the modelling of whole ensembles of models, i.e., 100 models of the same chromosomal region for both reference and SV-altered structures. My next step was to develop a tool that maps genomic features such as genes and enhancers, and calculates the spatial distance between pairs of gene promoters and enhancers. In the final step of this pipeline, these distributions are compared between the reference and the SV-altered variant. The results are displayed in a responsive table.

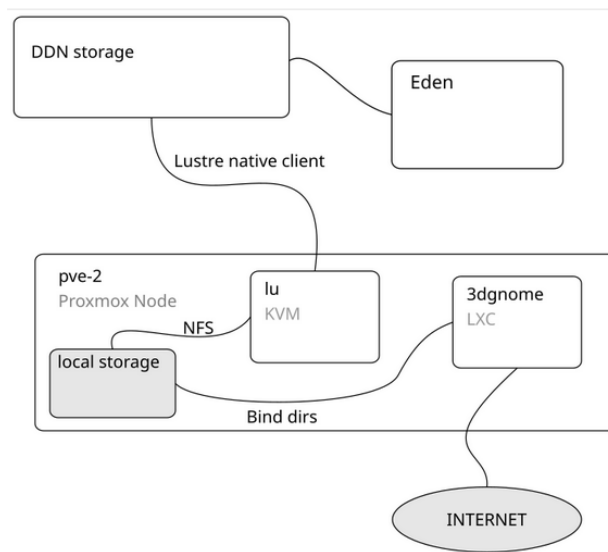


Figure 8 illustration the updated architecture of the 3D-GNOME web server that shows how cudaMMC integrates with the GPU-accelerated Eden cluster

Additionally, to enhance visualisation and data analysis, I integrated the IGV, an interactive genome browser, for viewing chromatin contact arcs along with additional annotations for genes, enhancers, and SVs. For 3D model visualisation, I also incorporated a new 3D viewer: NGL. Users can analyse the reference and SV-altered variants in two separate 3D viewers with labelled genes and enhancers mapped onto the 3D models, or they have the option to download them along with the rest of the results. The models are available for download in mmCIF and xyz formats.

Scientific achievements

Participation in research grants

- **[O1]** Three-dimensional Human Genome structure at the population scale: computational algorithm and experimental validation for lymphoblastoid cell lines of selected families from 1000 Genomes Project, institution: Centre of New Technologies, University of Warsaw, Poland, Awarding institution: FNP TEAM
- **[O2]** iCell: information processing in living organisms. The role of three-dimensional structure and multi-scale properties in controlling the biological processes in a cell institution: Centre of New Technologies, University of Warsaw, Poland
Awarding institution: NCN OPUS
- **[O3]** Enhpathy: Molecular Basis of Human enhanceropathies institution: Centre of New Technologies, University of Warsaw, Poland, Awarding institution: EU Horizon 2020 Marie Skłodowska-Curie
- **[O4]** BEYOND GWAS: tensor representation of context-specific regulatory variants for complex diseases and traits, institution: the Warsaw University of Technology, Poland, Awarding institution: IDUB PW

Academic visits

- **[V1]** Differences in genes activity between modern and archaic human populations mediated by changes in spatial chromatin structure, grant: MOBILITY PW, Visiting institution: University of Cambridge, under the supervision of Dr. Guy Jacobs,
- **[V2]** Investigation of differentiation of protein-mediated chromatin interactions on a population scale, including archaic human populations such as Neanderthals and

Denisovans , grant: RENOIR project (EU Horizon 2020 Marie Skłodowska-Curie)
Visiting institution: Complexity Institute at NTU, Singapore under the supervision
of Dr. Guy Jacobs,

- [V3] Identification of Epigenetic Biomarkers with the use of Gene Expression and DNA Methylation for Breast Cancer Subtypes
Visiting institution: National Institute of Technical Teachers' Training & Research (NITTTR), Kolkata, India under the supervision of Dr Indrajit Saha.

Conclusions and future work

With the rapid advancement of genomic technology, which continues to amass a wealth of data and genomic information processing within cell nuclei, there is a pressing need to develop computational tools and methods for processing, analysing, and decoding this information. One of the fields experiencing dynamic growth is the study of spatial chromatin organisation, particularly in terms of diversity across human populations, to investigate its impact on genomic activity.

This work aimed to address the need for computational tools to analyse the diversity of the 3D structure of the genome in different human populations, as well as to provide a synthetic approach for analysing transcription regulatory mechanisms, wherein regulation occurs through the 3D structure and alterations in distances between enhancers and promoters. As part of this work, we identified the necessity for high-performance IT tools for generating and analysing 3D chromatin structures on a population scale. I introduced a tool for predicting changes in chromatin contacts based on structural variants such as deletion, insertion, inversion, and duplication. Additionally, I presented the integration of this tool as a part of the 3D-GNOME 2.0 web server update. In conjunction with the integrated database of structural variants from the 1000 Genomes Project, users can analyse structural changes for selected human genomes. With the increase in data volumes and the demand for statistical analysis of entire 3D structure ensembles, we enhanced the 3D-GNOME tool for modelling 3D chromatin

structures by employing distributed calculations on GPU cards, which led to the development of the *cudaMMC* method. These tools enabled us to devise a method for analysing changes in the distribution of distances between significant enhancer-gene promoter pairs, which has been incorporated into the subsequent update of the 3D-GNOME web server.

The subsequent stages of development will involve applying these tools not only to population analysis but also to the examination of data from both healthy and diseased tissues. At present, tumour samples (e.g. breast cancer) are successfully analysed compared to healthy tissues, facilitating the investigation of potential biological mechanisms underlying tumour physiology. Nonetheless, integrating this data with genetic expression information continues to be a challenge for individual patients.

As experimental methods for obtaining data from archaic human populations, such as Neanderthals and Denisovans, advance, and with the emergence of a new field called paleo-informatics, there is a steadily increasing demand for the development of computational methods to process and analyse these data. This encompasses genetics, potential expression, comparison of chromatin 3D structures to those of modern human populations, and introgression between them. Key areas of focus include the development of identification techniques for structural variants, which is challenging due to the low quality of data coverage, and analysing data on a whole-population scale. Additionally, large-scale 3D modelling with progressively higher resolutions (down to the nucleosome level) is crucial, necessitating the development of more sophisticated 3D modelling tools. A limitation of this research is the still low quality of fossil data and their scarcity. Nonetheless, with ongoing discoveries and improvements in the detection of genomes and structural variants, research in this area is expected to progress significantly.

References

- [1] D. Malhotra and J. Sebat, “CNVs: harbingers of a rare variant revolution in psychiatric genetics,” *Cell*, vol. 148, no. 6, pp. 1223–1241, Mar. 2012, doi: 10.1016/j.cell.2012.02.039. [Online]. Available: <http://dx.doi.org/10.1016/j.cell.2012.02.039>
- [2] P. Stankiewicz and J. R. Lupski, “Structural variation in the human genome and its role in disease,” *Annu. Rev. Med.*, vol. 61, pp. 437–455, 2010, doi: 10.1146/annurev-med-100708-204735. [Online]. Available: <http://dx.doi.org/10.1146/annurev-med-100708-204735>
- [3] M. Zollino *et al.*, “Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype,” *Nat. Genet.*, vol. 44, no. 6, pp. 636–638, Apr. 2012, doi: 10.1038/ng.2257. [Online]. Available: <http://dx.doi.org/10.1038/ng.2257>
- [4] M. E. Talkowski *et al.*, “Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder,” *Am. J. Hum. Genet.*, vol. 89, no. 4, pp. 551–563, Oct. 2011, doi: 10.1016/j.ajhg.2011.09.011. [Online]. Available: <http://dx.doi.org/10.1016/j.ajhg.2011.09.011>
- [5] M. T. Maurano *et al.*, “Systematic localization of common disease-associated variation in regulatory DNA,” *Science*, vol. 337, no. 6099, pp. 1190–1195, Sep. 2012, doi: 10.1126/science.1222794. [Online]. Available: <http://dx.doi.org/10.1126/science.1222794>
- [6] P. H. Sudmant *et al.*, “An integrated map of structural variation in 2,504 human genomes,” *Nature*, vol. 526, no. 7571, pp. 75–81, Oct. 2015, doi: 10.1038/nature15394. [Online]. Available: <http://dx.doi.org/10.1038/nature15394>
- [7] D. G. Lupiáñez *et al.*, “Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions,” *Cell*, vol. 161, no. 5, pp. 1012–1025, May 2015, doi: 10.1016/j.cell.2015.04.004. [Online]. Available: <http://dx.doi.org/10.1016/j.cell.2015.04.004>
- [8] J. Weischenfeldt *et al.*, “Pan-cancer analysis of somatic copy-number alterations

- implicates IRS4 and IGF2 in enhancer hijacking,” *Nat. Genet.*, vol. 49, no. 1, pp. 65–74, Jan. 2017, doi: 10.1038/ng.3722. [Online]. Available: <http://dx.doi.org/10.1038/ng.3722>
- [9] P. A. Northcott *et al.*, “The whole-genome landscape of medulloblastoma subtypes,” *Nature*, vol. 547, no. 7663, pp. 311–317, Jul. 2017, doi: 10.1038/nature22973. [Online]. Available: <http://dx.doi.org/10.1038/nature22973>
- [10] J. R. Dixon *et al.*, “Integrative detection and analysis of structural variation in cancer genomes,” *Nat. Genet.*, vol. 50, p. 1388, 2018, doi: 10.1038/s41588-018-0195-8. [Online]. Available: <http://dx.doi.org/10.1038/s41588-018-0195-8>
- [11] D. Hnisz *et al.*, “Activation of proto-oncogenes by disruption of chromosome neighborhoods,” *Science*, vol. 351, pp. 1454–1458, 2016, doi: 10.1126/science.aad9024. [Online]. Available: <http://dx.doi.org/10.1126/science.aad9024>
- [12] S. Bianco *et al.*, “Polymer physics predicts the effects of structural variants on chromatin architecture,” *Nat. Genet.*, vol. 50, p. 662, 2018, doi: 10.1038/s41588-018-0098-8. [Online]. Available: <http://dx.doi.org/10.1038/s41588-018-0098-8>
- [13] M. Spielmann, D. G. Lupiáñez, and S. Mundlos, “Structural variation in the 3D genome,” *Nat. Rev. Genet.*, vol. 19, pp. 453–467, 2018, doi: 10.1038/s41576-018-0007-0. [Online]. Available: <http://dx.doi.org/10.1038/s41576-018-0007-0>
- [14] E. Lieberman-Aiden *et al.*, “Comprehensive mapping of long-range interactions reveals folding principles of the human genome,” *Science*, vol. 326, pp. 289–293, 2009, doi: 10.1126/science.1181369. [Online]. Available: <http://dx.doi.org/10.1126/science.1181369>
- [15] S. S. P. Rao *et al.*, “A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping,” *Cell*, vol. 159, pp. 1665–1680, 2014, doi: 10.1016/j.cell.2014.11.021. [Online]. Available: <http://dx.doi.org/10.1016/j.cell.2014.11.021>
- [16] M. J. Fullwood and Y. Ruan, “ChIP-based methods for the identification of long-range chromatin interactions,” *J. Cell. Biochem.*, vol. 107, pp. 30–39, 2009, doi: 10.1002/jcb.22116. [Online]. Available: <http://dx.doi.org/10.1002/jcb.22116>
- [17] Z. Tang *et al.*, “CTCF-mediated human 3D genome architecture reveals chromatin

- topology for transcription,” *Cell*, vol. 163, pp. 1611–1627, 2015, doi: 10.1016/j.cell.2015.11.024. [Online]. Available: <http://dx.doi.org/10.1016/j.cell.2015.11.024>
- [18] O. J. Luo, Z. Tang, X. Li, and Y. Ruan, “CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription.” 2015 [Online]. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72816>
- [19] C.-T. Ong and V. G. Corces, “CTCF: an architectural protein bridging genome topology and function,” *Nat. Rev. Genet.*, vol. 15, pp. 234–246, 2014, doi: 10.1038/nrg3663. [Online]. Available: <http://dx.doi.org/10.1038/nrg3663>
- [20] H. D. Ou, S. Phan, T. J. Deerinck, A. Thor, M. H. Ellisman, and C. C. O’Shea, “ChromEMT: visualizing 3D chromatin structure and compaction in interphase and mitotic cells,” *Science*. 2017 [Online]. Available: <https://science.sciencemag.org/node/697147.full>
- [21] S. S. P. Rao, M. H. Huntley, and E. Lieberman Aiden, “A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping.” 2014 [Online]. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525>
- [22] C. Chiang *et al.*, “The impact of structural variation on human gene expression,” *Nat. Genet.*, vol. 49, p. 692, 2017, doi: 10.1038/ng.3834. [Online]. Available: <http://dx.doi.org/10.1038/ng.3834>
- [23] M. Kasowski *et al.*, “Extensive variation in chromatin states across humans,” *Science*, vol. 342, pp. 750–752, 2013, doi: 10.1126/science.1242510. [Online]. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50893>
- [24] P. Szalaj *et al.*, “An integrated 3-Dimensional Genome Modeling Engine for data-driven simulation of spatial genome organization,” *Genome Res.*, vol. 26, pp. 1697–1709, 2016, doi: 10.1101/gr.205062.116. [Online]. Available: <http://dx.doi.org/10.1101/gr.205062.116>
- [25] W. de Laat and D. Duboule, “Topology of mammalian developmental enhancers and their regulatory landscapes,” *Nature*, vol. 502, pp. 499–506, 2013, doi: 10.1038/nature12753. [Online]. Available: <http://dx.doi.org/10.1038/nature12753>

- [26] “3D-GNOME 2.0 - 3D chromatin organization web service.” 2019 [Online]. Available: <https://3dgenome.cent.uw.edu.pl>
- [27] J. Ernst *et al.*, “Mapping and analysis of chromatin state dynamics in nine human cell types,” *Nature*, vol. 473, pp. 43–U52, 2011, doi: 10.1038/nature09906. [Online]. Available: <http://dx.doi.org/10.1038/nature09906>
- [28] J. M. Downen *et al.*, “Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes,” *Cell*, vol. 159, pp. 374–387, 2014, doi: 10.1016/j.cell.2014.09.030. [Online]. Available: <http://dx.doi.org/10.1016/j.cell.2014.09.030>
- [29] J. E. Phillips-Cremins *et al.*, “Architectural protein subclasses shape 3D Organization of Genomes during lineage commitment,” *Cell*, vol. 153, pp. 1281–1295, 2013, doi: 10.1016/j.cell.2013.04.053. [Online]. Available: <http://dx.doi.org/10.1016/j.cell.2013.04.053>
- [30] A. Buniello *et al.*, “The NHGRI-EBI GWAS Catalog of published genome-wide association studies (release 2018/01/31).” 2019.
- [31] B. Mifsud *et al.*, “Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C,” *Nat. Genet.*, vol. 47, pp. 598–606, 2015, doi: 10.1038/ng.3286. [Online]. Available: <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2323/>
- [32] P. Martin *et al.*, “Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci,” *Nat. Commun.*, vol. 6, 2015, doi: 10.1038/ncomms10069. [Online]. Available: <http://dx.doi.org/10.1038/ncomms10069>
- [33] D. J. Verlaan *et al.*, “Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease,” *Am. J. Hum. Genet.*, vol. 85, pp. 377–393, 2009, doi: 10.1016/j.ajhg.2009.08.007. [Online]. Available: <http://dx.doi.org/10.1016/j.ajhg.2009.08.007>
- [34] D. M. Altshuler *et al.*, “A global reference for human genetic variation,” *Nature*, vol. 526, p. 68, 2015, doi: 10.1038/nature15393. [Online]. Available: <http://dx.doi.org/10.1038/nature15393>

- [35] M. Y. Dennis and E. E. Eichler, “Human adaptation and evolution by segmental duplication,” *Curr. Opin. Genet. Dev.*, vol. 41, pp. 44–52, 2016, doi: 10.1016/j.gde.2016.08.001. [Online]. Available: <http://dx.doi.org/10.1016/j.gde.2016.08.001>
- [36] M. Y. Dennis *et al.*, “The evolution and population diversity of human-specific segmental duplications,” *Nature Ecology & Evolution*, vol. 1, 2017, doi: 10.1038/s41559-016-0069. [Online]. Available: <http://dx.doi.org/10.1038/s41559-016-0069>
- [37] A. H. Bittles, W. M. Mason, J. Greene, and N. A. Rao, “Reproductive-behavior and health in consanguineous marriages,” *Science*, vol. 252, pp. 789–794, 1991, doi: 10.1126/science.2028252. [Online]. Available: <http://dx.doi.org/10.1126/science.2028252>
- [38] D. Saleheen *et al.*, “Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity,” *Nature*, vol. 544, p. 235, 2017, doi: 10.1038/nature22034. [Online]. Available: <http://dx.doi.org/10.1038/nature22034>
- [39] T. Lappalainen *et al.*, “Transcriptome and genome sequencing uncovers functional variation in humans,” *Nature*, vol. 501, pp. 506–511, 2013, doi: 10.1038/nature12531. [Online]. Available: <http://dx.doi.org/10.1038/nature12531>
- [40] A. Schlattl, S. Anders, S. M. Waszak, W. Huber, and J. O. Korb, “Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions,” *Genome Res.*, vol. 21, pp. 2004–2013, 2011, doi: 10.1101/gr.122614.111. [Online]. Available: <http://dx.doi.org/10.1101/gr.122614.111>
- [41] B. E. Stranger *et al.*, “Relative impact of nucleotide and copy number variation on gene expression phenotypes,” *Science*, vol. 315, pp. 848–853, 2007, doi: 10.1126/science.1136678. [Online]. Available: <http://dx.doi.org/10.1126/science.1136678>
- [42] J. K. Pickrell *et al.*, “Understanding mechanisms underlying human gene expression variation with RNA sequencing,” *Nature*, vol. 464, pp. 768–772, 2010, doi: 10.1038/nature08872. [Online]. Available: <http://dx.doi.org/10.1038/nature08872>
- [43] J.-B. Veyrieras *et al.*, “High-resolution mapping of expression-QTLs yields insight into

- human gene regulation,” *PLoS Genet.*, vol. 4, no. 10, p. e1000214, Oct. 2008, doi: 10.1371/journal.pgen.1000214. [Online]. Available: <http://dx.doi.org/10.1371/journal.pgen.1000214>
- [44] D. J. Gaffney *et al.*, “Dissecting the regulatory architecture of gene expression QTLs,” *Genome Biol.*, vol. 13, no. 1, p. R7, 2012, doi: 10.1186/gb-2012-13-1-r7. [Online]. Available: <http://dx.doi.org/10.1186/gb-2012-13-1-r7>
- [45] E. Eisenberg and E. Y. Levanon, “Human housekeeping genes, revisited,” *Trends Genet.*, vol. 30, pp. 119–119, 2014, doi: 10.1016/j.tig.2014.02.001. [Online]. Available: <http://dx.doi.org/10.1016/j.tig.2014.02.001>
- [46] J. R. Dixon *et al.*, “Topological domains in mammalian genomes identified by analysis of chromatin interactions,” *Nature*, vol. 485, pp. 376–380, 2012, doi: 10.1038/nature11082. [Online]. Available: <http://dx.doi.org/10.1038/nature11082>
- [47] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, “Capturing chromosome conformation,” *Science*, vol. 295, pp. 1306–1311, 2002, doi: 10.1126/science.1067799. [Online]. Available: <http://dx.doi.org/10.1126/science.1067799>
- [48] I. Dunham *et al.*, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, pp. 57–74, 2012, doi: 10.1038/nature11247. [Online]. Available: <http://dx.doi.org/10.1038/nature11247>
- [49] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, pp. 841–842, 2010, doi: 10.1093/bioinformatics/btq033. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btq033>
- [50] D. E. Schones, A. D. Smith, and M. Q. Zhang, “Statistical significance of cis-regulatory modules,” *BMC Bioinformatics*, vol. 8, p. 19, 2007, doi: 10.1186/1471-2105-8-19. [Online]. Available: <http://dx.doi.org/10.1186/1471-2105-8-19>
- [51] A. Khan *et al.*, “JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework,” *Nucleic Acids Res.*, vol. 46, pp. D260–D266, 2018, doi: 10.1093/nar/gkx1126. [Online]. Available: <http://dx.doi.org/10.1093/nar/gkx1126>

- [52] J. Ernst and M. Kellis, “ChromHMM: automating chromatin-state discovery and characterization,” *Nat. Methods*, vol. 9, pp. 215–216, 2012, doi: 10.1038/nmeth.1906. [Online]. Available: <http://dx.doi.org/10.1038/nmeth.1906>
- [53] A. Frankish *et al.*, “GENCODE reference annotation for the human and mouse genomes,” 2019. [Online]. Available: ftp://ftp.ebi.ac.uk/pub/databases/genencode/Gencode_human/
- [54] O. Wagih, “ggseqlogo: a versatile R package for drawing sequence logos,” *Bioinformatics*, vol. 33, pp. 3645–3647, 2017, doi: 10.1093/bioinformatics/btx469. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btx469>
- [55] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin, “Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses,” *Nat. Protoc.*, vol. 7, pp. 500–507, 2012, doi: 10.1038/nprot.2011.457. [Online]. Available: <http://dx.doi.org/10.1038/nprot.2011.457>
- [56] K. Breuer *et al.*, “InnateDB: systems biology of innate immunity and beyond - recent updates and continuing curation,” 2013. [Online]. Available: <https://www.innatedb.com/annotatedGenes.do?type=innatedb>
- [57] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nat. Methods*, vol. 9, pp. 357–U354, 2012, doi: 10.1038/nmeth.1923. [Online]. Available: <http://dx.doi.org/10.1038/nmeth.1923>
- [58] F. Ramirez *et al.*, “deepTools2: a next generation web server for deep-sequencing data analysis,” *Nucleic Acids Res.*, vol. 44, pp. W160–W165, 2016, doi: 10.1093/nar/gkw257. [Online]. Available: <http://dx.doi.org/10.1093/nar/gkw257>
- [59] D. R. Zerbino, N. Johnson, T. Juettemann, S. P. Wilder, and P. Flicek, “WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis,” *Bioinformatics*, vol. 30, pp. 1008–1009, 2014, doi: 10.1093/bioinformatics/btt737. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btt737>
- [60] J. Rozowsky *et al.*, “AlleleSeq: analysis of allele-specific expression and binding in a network framework,” *Mol. Syst. Biol.*, vol. 7, p. 522, 2011, doi: 10.1038/msb.2011.54. [Online]. Available: <http://dx.doi.org/10.1038/msb.2011.54>

- [61] H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, and L. Wang, “CrossMap: a versatile tool for coordinate conversion between genome assemblies,” *Bioinformatics*, vol. 30, pp. 1006–1007, 2014, doi: 10.1093/bioinformatics/btt730. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btt730>
- [62] “Minimal Steps For LiftOver,” 2019. [Online]. Available: http://genomewiki.ucsc.edu/index.php/Minimal_Steps_For_LiftOver
- [63] P. Danecek *et al.*, “The variant call format and VCFtools,” *Bioinformatics*, vol. 27, pp. 2156–2158, 2011, doi: 10.1093/bioinformatics/btr330. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btr330>
- [64] S. Purcell *et al.*, “PLINK: a tool set for whole-genome association and population-based linkage analyses,” *Am. J. Hum. Genet.*, vol. 81, pp. 559–575, 2007, doi: 10.1086/519795. [Online]. Available: <http://dx.doi.org/10.1086/519795>
- [65] D. R. Zerbino *et al.*, “Ensembl 2018,” *Nucleic Acids Res.*, vol. 46, pp. D754–D761, 2018, doi: 10.1093/nar/gkx1098. [Online]. Available: <http://dx.doi.org/10.1093/nar/gkx1098>
- [66] A. S. Hinrichs *et al.*, “The UCSC Genome Browser Database: update 2006,” *Nucleic Acids Res.*, vol. 34, pp. D590–D598, 2006, doi: 10.1093/nar/gkj144. [Online]. Available: <http://dx.doi.org/10.1093/nar/gkj144>
- [67] M. Sadowski, “Spatial chromatin architecture alteration by structural variations in human genomes at the population scale [code].” 2019.
- [68] S. Schoenfelder and P. Fraser, “Long-range enhancer-promoter contacts in gene expression control,” *Nat. Rev. Genet.*, vol. 20, pp. 437–455, 2019.
- [69] M. Sadowski *et al.*, “Spatial chromatin architecture alteration by structural variations in human genomes at the population scale,” *Genome Biol.*, vol. 20, no. 1, p. 148, Jul. 2019, doi: 10.1186/s13059-019-1728-x. [Online]. Available: <http://dx.doi.org/10.1186/s13059-019-1728-x>
- [70] A. Auton *et al.*, “A global reference for human genetic variation,” *Nature*, vol. 526, pp. 68–74, 2015.
- [71] S. Heinz *et al.*, “Transcription Elongation Can Affect Genome 3D Structure,” *Cell*, vol.

- 174, pp. 1522–1536 e1522, 2018.
- [72] O. L. Kantidze, K. V. Gurova, V. M. Studitsky, and S. V. Razin, “The 3D Genome as a Target for Anticancer Therapy,” *Trends Mol. Med.*, vol. 26, pp. 141–149, 2020.
- [73] S. Mallick *et al.*, “The Simons Genome Diversity Project: 300 genomes from 142 diverse populations,” *Nature*, vol. 538, pp. 201–206, 2016.
- [74] P. Szalaj *et al.*, “3D-GNOME: an integrated web service for structural modeling of the 3D genome,” *Nucleic Acids Res.*, vol. 44, pp. W288–W293, 2016.
- [75] M. J. P. Chaisson *et al.*, “Multi-platform discovery of haplotype-resolved structural variation in human genomes,” *Nat. Commun.*, vol. 10, p. 1784, 2019.
- [76] M. Chiliński, K. Sengupta, and D. Plewczynski, “From DNA human sequence to the chromatin higher order organisation and its biological meaning: Using biomolecular interaction networks to understand the influence of structural variation on spatial genome organisation and its functional effect,” in *Seminars in Cell & Developmental Biology*, Aug. 2021.
- [77] J. Ray *et al.*, “Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 39, pp. 19431–19439, 2019.
- [78] L. Pei *et al.*, “Dynamic 3D genome architecture of cotton fiber reveals subgenome-coordinated chromatin topology for 4-staged single-cell differentiation,” *Genome Biol.*, vol. 23, no. 1, pp. 1–25, 2022.
- [79] M. Lazniewski, W. K. Dawson, A. M. Rusek, and D. Plewczynski, “One protein to rule them all: the role of CCCTC-binding factor in shaping human genome in health and disease,” in *Seminars in cell & developmental biology*, Jun. 2019, vol. 90, pp. 114–127.
- [80] M. Kadlof, J. Rozycka, and D. Plewczynski, “Spring Model–chromatin modeling tool based on OpenMM,” *Methods*, vol. 181, pp. 62–69, 2020.
- [81] P. Szałaj *et al.*, “An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization,” *Genome Res.*, vol. 26, no. 12, pp. 1697–1709, 2016.

- [82] M. Di Pierro, B. Zhang, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, “Transferable model for chromosome architecture,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 43, pp. 12168–12173, 2016.
- [83] M. Wlasnowolski *et al.*, “3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome,” *Nucleic Acids Res.*, vol. 48, no. W1, pp. W170–W176, 2020.
- [84] E. F. Pettersen *et al.*, “UCSF Chimera—a visualization system for exploratory research and analysis,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.
- [85] L. Isbel, R. S. Grand, and D. Schübeler, “Generating specificity in genome regulation through transcription factor sensitivity to chromatin,” *Nat. Rev. Genet.*, vol. 23, no. 12, pp. 728–740, 2022.
- [86] A. Hafner and A. Boettiger, “The spatial organization of transcriptional control,” *Nat. Rev. Genet.*, pp. 1–16, 2022.
- [87] P. J. Farnham, “Insights from genomic profiling of transcription factors,” *Nat. Rev. Genet.*, vol. 10, no. 9, pp. 605–616, 2009.
- [88] N. D. Heintzman and B. Ren, “Finding distal regulatory elements in the human genome,” *Curr. Opin. Genet. Dev.*, vol. 19, no. 6, pp. 541–549, 2009.
- [89] M. Levine, “Transcriptional enhancers in animal development and evolution,” *Curr. Biol.*, vol. 20, no. 17, pp. R754–R763, 2010.
- [90] A. J. Scott and C. C. H.i., “Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes,” *Genome Res.*, vol. 526, no. 12, pp. 2249–2257, 2021.
- [91] A. Gusev *et al.*, “Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases,” *Am. J. Hum. Genet.*, vol. 95, no. 5, pp. 535–552, 2014.
- [92] Y. Liu *et al.*, “Application of deep learning algorithm on whole genome sequencing data uncovers structural variants associated with multiple mental disorders in african american patients,” *Mol. Psychiatry*, vol. 27, no. 3, pp. 1469–1478, 2022.
- [93] Z. Liu, R. Roberts, T. R. Mercer, J. Xu, F. J. Sedlazeck, and W. Tong, “Towards

- accurate and reliable resolution of structural variants for clinical diagnosis,” *Genome Biol.*, vol. 23, no. 1, p. 68, 2022.
- [94] M. Wlasnowolski *et al.*, “3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome,” *Nucleic Acids Res.*, vol. 48, no. W1, pp. W170–W176, 2020.
- [95] T. I. G. P. Consortium, “A global reference for human genetic variation,” *Journal of Open Source Software*, vol. 526, no. 60, pp. 68–74, 2015.
- [96] P. Szałaj *et al.*, “An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization,” *Genome Res.*, vol. 26, no. 12, pp. 1697–1709, 2016.
- [97] A. S. Rose and P. W. Hildebrand, “NGL Viewer: a web application for molecular visualization,” *Nucleic Acids Res.*, vol. 43, no. W1, pp. W576–W579, 2015.
- [98] M. Kadlof, J. Rozycka, and D. Plewczynski, “Spring Model--chromatin modeling tool based on OpenMM,” *Methods*, vol. 181, pp. 62–69, 2020.
- [99] Y. Wang *et al.*, “The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions,” *Genome Biol.*, vol. 19, no. 1, pp. 1–12, 2018.
- [100] X. Li *et al.*, “Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions,” *Nat. Protoc.*, vol. 12, no. 5, pp. 899–915, 2017.
- [101] P. Wang *et al.*, “In situ chromatin interaction analysis using paired-end tag sequencing,” *Nat. Protoc.*, vol. 16, no. 3, pp. 1489–1519, 2021.
- [102] N. Desai and W. Cirne, *Job Scheduling Strategies for Parallel Processing: 17th International Workshop, JSSPP 2013, Boston, MA, USA, May 24, 2013 Revised Selected Papers*. Springer, 2014 [Online]. Available: <https://play.google.com/store/books/details?id=S2e5BQAAQBAJ>
- [103] O. Tange, *GNU Parallel 2018*. Lulu.com, 2018 [Online]. Available: https://books.google.com/books/about/GNU_Parallel_2018.html?hl=&id=sKdSDwAAQBAJ

Copies of the publications constituting the Dissertation


- [P1] Sadowski, M., Kraft, A., Szalaj, P., **Wlasnowolski, M.**, Tang, Z., Ruan, Y., & Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome biology*, 20(1), 1-27.
IF=18.01, MNiSW points: 200
- [P2] **Wlasnowolski, M.**, Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y. & Plewczynski, D. (2020). 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Research*, [online] 48(W1), pp.W170–W176.
IF=19.16, MNiSW points: 200, related to ITT discipline
- [P3] **Wlasnowolski, M.**, Grabowski, P., Roszczyk, D., Kaczmarek, K., & Plewczynski, D. (2023). cudaMMC - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling. bioRxiv. <https://doi.org/10.1101/2023.06.12.544609>
in revision in *Bioinformatics* (IF=6.931, MNiSW points: 200, related to ITT discipline)
- [P4] **Wlasnowolski, M.**, Kadlof, M., Sengupta, K., & Plewczynski, D. (2023). 3D-GNOME 3.0: a three-dimensional genome modelling engine for analysing changes of promoter-enhancer contacts in the human genome. *Nucleic Acids Research*, gkad354.
IF=19.16, MNiSW points: 200, related to ITT discipline

RESEARCH

Open Access



Spatial chromatin architecture alteration by structural variations in human genomes at the population scale

Michał Sadowski^{1,2}, Agnieszka Kraft^{1,3}, Przemysław Szalaj^{1,4,5}, Michał Wlasnowolski^{1,3}, Zhonghui Tang⁶, Yijun Ruan^{7*} and Dariusz Plewczynski^{1,3*} 

Abstract

Background: The number of reported examples of chromatin architecture alterations involved in the regulation of gene transcription and in disease is increasing. However, no genome-wide testing has been performed to assess the abundance of these events and their importance relative to other factors affecting genome regulation. This is particularly interesting given that a vast majority of genetic variations identified in association studies are located outside coding sequences. This study attempts to address this lack by analyzing the impact on chromatin spatial organization of genetic variants identified in individuals from 26 human populations and in genome-wide association studies.

Results: We assess the tendency of structural variants to accumulate in spatially interacting genomic segments and design an algorithm to model chromatin conformational changes caused by structural variations. We show that differential gene transcription is closely linked to the variation in chromatin interaction networks mediated by RNA polymerase II. We also demonstrate that CTCF-mediated interactions are well conserved across populations, but enriched with disease-associated SNPs. Moreover, we find boundaries of topological domains as relatively frequent targets of duplications, which suggest that these duplications can be an important evolutionary mechanism of genome spatial organization.

Conclusions: This study assesses the critical impact of genetic variants on the higher-order organization of chromatin folding and provides insight into the mechanisms regulating gene transcription at the population scale, of which local arrangement of chromatin loops seems to be the most significant. It provides the first insight into the variability of the human 3D genome at the population scale.

Keywords: Genomics, Chromatin architecture, Topologically associating domains, Chromatin loops, Genome regulation, Gene transcription, CCCTC-binding factor, RNA polymerase II, Biophysical modeling, Human

Background

Around 20 million base pairs of a normal human genome (0.6%) are under structural variations, including deletions, duplications, insertions, and inversions. This makes structural variants (SVs) the most prominent source of genetic variation among human individual genomes.

The potential malicious effect of SVs has been recognized but almost solely associated with altering gene copy number and gene structure—a number of studies relate copy number variants (CNVs) affecting gene regions to cancer [1], intellectual disabilities [2], and predispositions to various health problems [3, 4]. The vast majority of genetic variation occurs, however, in non-coding regions. Over 95% of single-nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS) are located outside coding sequences [5]. Similarly, larger variants are significantly depleted in gene regions [6].

* Correspondence: Yijun.Ruan@jax.org; d.plewczynski@cent.uw.edu.pl

⁷The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, USA

¹Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland

Full list of author information is available at the end of the article



A part of the SVs emerging in non-coding regions alters genomic loci recognized by proteins which organize the human genome in the cell nuclear space. Recent studies provided some insights into the impact SVs can have on a spatial organization of the human genome. Examples of SVs altering the borders of TADs in *EPHA4* locus and causing pathogenic phenotypes by enabling spatial contacts between formerly isolated genomic functional elements were reported [7]. Positions of TAD boundaries were proven useful for inferring cancer-related gene overexpression resulting from variation in *cis*-regulatory elements [8]. Accumulation of SVs proximal to the TAD boundary occupied by CTCF was postulated to cause enhancer hijacking and *PRDM6* overexpression in medulloblastoma samples [9]. Hi-C maps were successfully used for the detection of large-scale rearrangements, which were reported as frequent in cancer cells [10]. Disruptions of chromosome neighborhoods were demonstrated—using CRISPR/Cas9 experiments—to activate proto-oncogenes [11]. An attempt was also made to model 3D chromatin structure including information on SVs and predicting enrichment/depletion of higher-order chromatin contacts caused by these variations [12]. Efficacy of the modeling method in predicting SV-induced ectopic contacts at the level of TADs was shown for *EPHA4* locus.

However, to our knowledge, there was no genome-wide systematic study on the impact of SVs on genome spatial organization analyzing the level of individual chromatin loops. One of the most recent reviews on the topic [13] highlights the impact of SVs on genome spatial structure and the pathogenic potential of SVs altering the higher-order chromatin organization. Nonetheless, no attempt was made by the authors to assess what part of SVs emerging in normal human genomes causes functionally relevant chromatin spatial rearrangements, and no genome-wide data was presented on how SVs influence the chromatin 3D architecture.

The recent advancements in chromosome conformation capture techniques, namely high-throughput conformation capture (Hi-C) [14, 15] and Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) [16, 17], resulted in the release of high-resolution chromatin interaction datasets. ChIA-PET, in particular, is able to capture individual chromatin contacts mediated by specific protein factors. In turn, the great effort of the 1000 Genomes Consortium led to the creation of the catalog of human genomic sequence variations [6] identified in over 2500 human samples from 26 populations.

Taking advantage of the high-quality ChIA-PET and population-scale SVs data, we discuss a mechanistic model of the impact of SVs on the chromatin looping structure, provide the first genome-wide analysis of this impact for the human genome, and

model SV-induced changes in 3D genomic structures observed in human population.

In our analyses of the impact of SVs on the 3D chromatin organization of the human genome, we pay a specific attention to chromatin interactions associated with enhancer regions and gene promoters. These interactions are likely to play a distinguished role in the regulation of gene transcription in a mechanistic fashion, bringing the genes and the regulatory elements close together or separating them in the nuclear space of the cell. We observe an interesting interplay between such genomic interactions and SVs.

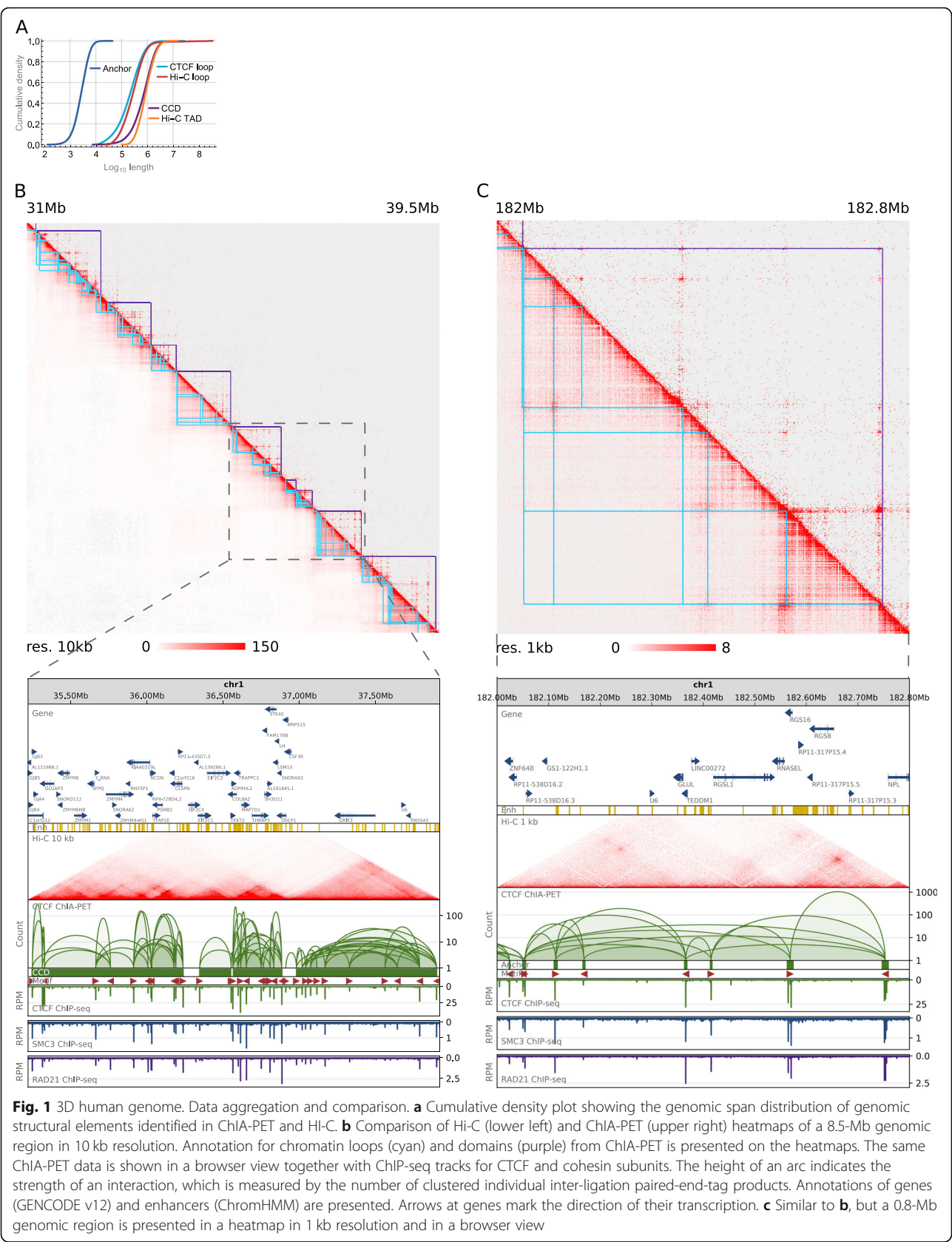
Results

3D human genome

In this study, we use ChIA-PET interactions as a representation of the higher-order spatial organization of the human genome. ChIA-PET targeting on CTCF and RNAPII performed on the GM12878 cell line [17, 18] was selected as the most comprehensive ChIA-PET dataset for humans presently. CTCF was shown to be the key protein factor shaping the architecture of mammalian genomes [15, 19], whereas RNAPII is essential for gene transcription. Together, the ChIA-PET data of these two protein factors account for structural and functional aspects of the higher-order organization and multiscale folding of chromatin 10 nm fiber in the human cell nucleus [20]. It was postulated that pools of interacting CTCF/cohesin-mediated loop anchors form the structural foci, toward which interactions mediated by RNAPII draw genes for coordinated transcription [17]. We will further refer to these structural foci as interaction centers.

ChIA-PET generates high-resolution (~ 100 bp) genome-wide chromatin contact maps. It identifies two types of chromatin interactions mediated by specific protein factors. The first type is highly reliable enriched interactions which appear in the data as closely mapped on the genome clustered inter-ligation paired-end-tag products (PET clusters). The second type is singletons, which reflect higher-order topological proximity [17].

We inspected the anchoring sites of PET clusters identified by the CTCF ChIA-PET experiment for the co-occupancy by CTCF and cohesin (SMC3 and RAD21 subunits), to select the set of high-quality chromatin interactions mediated by CTCF in GM12878 cell (see the “Methods” section). We identified 44,380 such pairwise interactions (Additional file 1: Table S1). The median length of genomic segments joined by these interactions is 2730 bp, and 99% of them are shorter than 10 kb (Fig. 1a). Nucleotide sequences of these segments usually contain multiple CTCF motifs. Chromatin loops formed by the CTCF interactions have lengths in the order of 100 kb (Fig. 1a).



The interactions mediated by CTCF are not uniformly distributed over the genome but rather form highly interacting, predominantly hundreds of kilobases-long chromatin blocks (which we will further refer to as chromatin contact domains (CCDs)) separated by segments of weak and rare contacts (gaps). Based on the CTCF ChIA-PET data, the genome of GM12878 cell was segmented into 2267 CCDs [18], and we adopt this segmentation in this study. The domains lengths vary from around 10 kb to few megabases with a median length of 750 kb. Only 1% of CCDs is longer than 2 Mb (Fig. 1a).

Even though CTCF ChIA-PET captures only the interactions mediated by the CTCF protein, it detects structural features exhibited by the non-specific Hi-C data [21]. It was shown that at a global scale, whole-chromosome Hi-C and ChIA-PET contact maps are highly correlated (Spearman's correlation coefficient in the range of 0.7–0.9) [17]. Locally, ChIA-PET and Hi-C heatmaps identify a very similar landscape of genomic structures, both at the scale of topological domains (Fig. 1b) and chromatin loops (Fig. 1c). A large fraction of TADs identified in high-resolution Hi-C data are demarcated by anchors of chromatin loops highly enriched with CTCF and the cohesin subunits SMC3 and RAD21 [15]. Even though the borders of topological domains formed by CTCF loops identified in ChIA-PET data coincide with only a small fraction of anchors of those loops, they exhibit distinctively high levels of CTCF, SMC3, and RAD21 binding signals (Fig. 1b and Additional file 2: Figure S1). This underlines the specificity of those loci among CTCF loop anchors and is consistent with the findings based on Hi-C data. Furthermore, length distributions of chromatin loops and topological domains called from ChIA-PET data are concordant with the respective statistics for Hi-C (Fig. 1a). All this indicates that CTCF ChIA-PET dataset generated for GM12878 cell is a high-quality representation of the human 3D genome.

Following the authors of the dataset, we investigated the directionality of CTCF motifs in the anchors of the CTCF chromatin loops. Thirty-seven thousand two hundred eighty-nine out of the 44,380 PET clusters had motifs of unique orientation in both anchors. Among the 37,289, we found 24,181 (65%) interactions with motifs in the anchors having convergent orientation (convergent loops), 6118 (16%) interactions in tandem right orientation (tandem right loops), 6089 (16%) tandem left loops, and 901 (2%) divergent loops (see the “Methods” section). We adopted the coordinates of the outermost CTCF motifs in CCDs as indicators of their borders (see the “Methods” section).

The described ChIA-PET dataset is further used as the reference 3D genome of a human lymphoblastoid cell.

Predicting chromatin architecture altered by SVs

It was demonstrated, by phasing CTCF PET clusters identified in GM12878, that allele-specific single-nucleotide variation in genome sequence can result in haplotype-specific chromatin topology [17].

We further show that relative values of haplotype-specific CTCF binding signals (see the “Methods” section) accurately reflect genotypes determined by this variation in a number of lymphoblastoid cell lines (Fig. 2a, b; Additional file 2: Figure S2A and S2B). Furthermore, CTCF binding profiles around CTCF interaction anchors of unchanged nucleotide sequences are very similar across the cells. The analogy between the changes in CTCF binding caused by anchor-targeting allele-specific SNPs between homologous chromosomes of GM12878 and among lymphoblastoid cells suggests that the major differences in chromatin topology between chromosomes of two lymphoblastoid cells are an effect of genetic variation and can be predicted based on genomic interactions identified in GM12878. Such predictions can in turn uncover causal relations underlying the associations observed between genetic variations and gene transcription rates (Fig. 2c).

In this study, we concentrate on predicting how SVs impact genome looping organization. SVs are the major source of sequence variation among human genomes and given their larger size have a higher potential than SNPs to induce changes in chromatin folding. They were also shown to contribute more than SNPs to variation in gene expression among human samples [22]. Chromatin contacts are thought to be largely invariant across individuals. To assess the level of conservation of CTCF-dependent genome architecture across individuals, we analyzed the abundance and arrangement of CTCF ChIP-seq peaks from 13 lymphoblastoid samples in genomic segments which were identified as CTCF-mediated interaction anchors in the GM12878 cell line. The ChIP-seq data originate from one study [23, 24] and include 5 samples of European ancestry (GM10847, GM12878, GM12890, GM12891, GM12892), 7 samples of African ancestry (GM18486, GM18505, GM19099, GM19193, GM19238, GM19239, GM19240), and 2 samples of East Asian ancestry (GM18526, GM18951). GM12891, GM12892, GM12878 and GM19239, GM19238 and GM19240 are families of father, mother, and child respectively. The datasets were rigorously filtered for comparisons (see the “Methods” section).

Our analysis shows that CTCF binding profiles at long-range interaction sites are highly similar across lymphoblastoid cells. Over 99% of interacting anchors occupied by CTCF peaks in GM12878 cell are supported by CTCF peaks in each of the 13 other samples (Fig. 2d). Moreover, the overall distribution of CTCFs involved in the formation of chromatin contacts is shared among individuals.



(See figure on previous page.)

Fig. 2 Predicting the impact of SVs on the chromatin topology. **a** Browser view of a 0.5-Mb genomic segment with asthma-associated SNP rs12936231 identified in a part of the human population. SNP rs12936231 alters the sequence of a CTCF motif involved in interactions. Haplotype-specific CTCF signals from 10 lymphoblastoid cells are presented along with haplotype-specific CTCF ChIA-PET interactions from GM12878 (only a subset of all interactions can be identified as specifically paternal/maternal as it is done based on allele-specific SNPs emerging at the interaction anchors). For each track, ChIP-seq signal values (originally in RPMs) were divided by the maximal value of the signal in the visualized region. Sum of the signal values over the genomic region occupied by the SNP-affected interaction anchor together with the genotype is marked in each signal track. **b** Comparison of sequences and scores of CTCF binding motifs carrying the reference C and the alternative G alleles of rs12936231. **c** Differences in gene transcription rates between genotypes set for rs12936231. Genes exhibiting differences in transcription which pass Mann-Whitney test with p value < 0.05 were reported. Center lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range (IQR) from the 25th and 75th percentiles; outliers are represented by rings; far outliers (points beyond 3 times the IQR) are not represented by any element of box plots. $n = 101, 227, 117$ sample points. **d** CTCF anchors from GM12878 not intersected with CTCF ChIP-seq peaks identified in different lymphoblastoid cells. The anchors were filtered by consensus CTCF binding sites (see the “Methods” section). **e** Number of SVs, divided by type, intersecting (in case of interaction anchors), covering (in case of CCD boundaries), or contained in (in case of CCDs and CCD gaps) different genomic structural elements

For about 90% of all CTCF peaks identified in anchors of PET clusters in GM12878 cell, a CTCF peak in each of the compared cells can be found within a distance of 400 bp (Additional file 2: Figure S3). CTCF-interacting anchors and borders of CCDs identified in GM12878 cell are highly enriched with CTCF ChIP-seq peaks found in the other 13 lymphoblastoid cells (Additional file 2: Figure S4).

Similar analyses were performed for RNAPII-mediated interacting anchors using RNAPII ChIP-seq peaks (Additional file 2: Figure S5). Significantly bigger but moderate differences among samples were observed for this data.

Having tested the resemblance of genomic interaction site distribution in lymphoblastoid cells, we match SVs detected in human population in the 1000 Genomes Project [6] with the reference network of chromatin interactions to obtain individualized chromatin interaction patterns and assess the topological variability among human genomes.

There are 68,818 unique SVs deposited in the 1000 Genomes Catalog of Human Genetic Variation (CHGV) [25], including deletions, duplications, multi-allelic CNVs (mCNVs), inversions, and insertions (Fig. 2e). Forty-four percent of them are shorter than 1 kb, and only 22% is longer than 10 kb (Additional file 2: Figure S6).

Most of the SVs reside inside CCDs, not intersecting borders of domains nor CTCF-mediated interaction anchors (Fig. 2e).

Computational algorithm for modeling SV-induced chromatin conformational changes

While the SVs that miss the interacting binding sites in most cases have limited impact on the final structure (resulting only in shortening, or extending of the corresponding chromatin loops), the SVs that overlap the interacting sites may partially modify the interaction pattern and in turn cause serious changes of the 3D structure (Fig. 3a). Specifically, deletion removes an

interacting anchor therefore deleting all chromatin loops mediated by this genomic site; duplication introduces a new interaction site, which has the same underlying sequence specificity to form chromatin loops as the original duplicated site; inversion which encircles a CTCF binding motif will revert its directionality therefore affecting the chromatin looping of the neighboring region; and finally, insertion containing CTCF motif enables new interaction sites capable of forming chromatin loops with other CTCF binding sites.

We earlier demonstrated that using CTCF ChIA-PET data, a 3D model of an averaged genome structure which recovers architectural features of the genome can be built [26]. Chromatin models constructed with our computational tool (3D-GNOME) can be used to illustrate the most probable arrangement of genomic structural elements in 3D space: from chromosomes, through topological domains, to individual chromatin loops (see the “Methods” section). 3D-GNOME uses PET clusters to position the binding sites relative to each other first and then employs singletons, orientations of the CTCF motifs, and biophysical constraints to accurately model the shape of individual chromatin loops (Fig. 3b).

We extended the 3D-GNOME modeling approach to include information on SVs in the recovery of 3D chromatin structures (see the “Methods” section). Our algorithm models individual chromatin loops, meaning that the remodeling effect of a genetic variant disrupting a single pair of interacting genomic segments will be represented in the model (Fig. 3a). 3D-GNOME is an optimization algorithm which returns models that fulfill spatial constraints coming from genomic interaction data. Typically, a number of solutions exist for a given set of constraints (Fig. 3a).

Analysis and modeling of genome organization levels of topological domains and chromatin loops are in the main scope of this study, as topological domains are believed to be the structural units regulating gene transcription by spatially isolating groups of enhancers

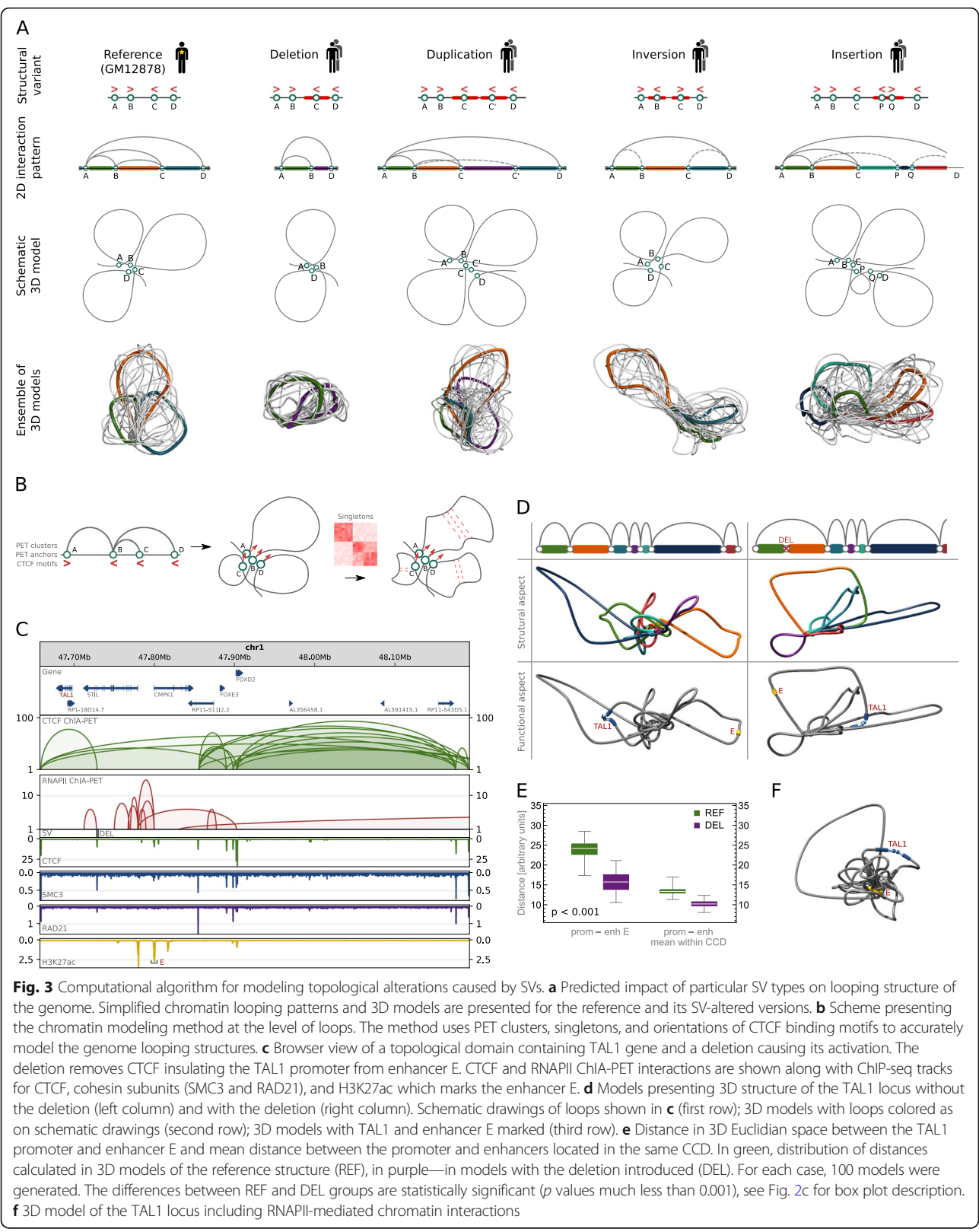


Fig. 3 Computational algorithm for modeling topological alterations caused by SVs. **a** Predicted impact of particular SV types on looping structure of the genome. Simplified chromatin looping patterns and 3D models are presented for the reference and its SV-altered versions. **b** Scheme presenting the chromatin modeling method at the level of loops. The method uses PET clusters, singletons, and orientations of CTCF binding motifs to accurately model the genome looping structures. **c** Browser view of a topological domain containing TAL1 gene and a deletion causing its activation. The deletion removes CTCF insulating the TAL1 promoter from enhancer E. CTCF and RNAPII ChIA-PET interactions are shown along with ChIP-seq tracks for CTCF, cohesin subunits (SMC3 and RAD21), and H3K27ac which marks the enhancer E. **d** Models presenting 3D structure of the TAL1 locus without the deletion (left column) and with the deletion (right column). Schematic drawings of loops shown in **c** (first row); 3D models with loops colored as on schematic drawings (second row); 3D models with TAL1 and enhancer E marked (third row). **e** Distance in 3D Euclidian space between the TAL1 promoter and enhancer E and mean distance between the promoter and enhancers located in the same CCD. In green, distribution of distances calculated in 3D models of the reference structure (REF), in purple—in models with the deletion introduced (DEL). For each case, 100 models were generated. The differences between REF and DEL groups are statistically significant (p values much less than 0.001), see Fig. 2c for box plot description. **f** 3D model of the TAL1 locus including RNAPII-mediated chromatin interactions

and genes [7, 15, 17, 27]. In our opinion, our 3D models constitute a supportive insight into SV effects; their inspection can improve the understanding of functional impact and disease association of SVs.

As an example, deletion of a CTCF binding site insulating the promoter of TAL1 gene from regulatory elements adjacent to the CMPK1 promoter was shown by CRISPR/Cas9 experiments to cause activation of TAL1, an oncogenic driver of T cell acute lymphoblastic leukemia [11] (Fig. 3c). 3D structures of the TAL1 locus generated with our algorithm illustrate fusion of the TAL1 promoter with the enhancer regions inside the insulated neighborhood formed as a consequence of the deletion (Fig. 3d). 3D distances calculated from the models quantify the accessibility of transcription-enhancing elements for the TAL1 promoter. The 3D distance between the promoter and a strong enhancer in the CMPK1 promoter and the mean distance between the promoter and enhancers located in the same CCD decrease significantly after the deletion (Fig. 3e). The models show how the promoter and the enhancer are brought even closer together within the insulated neighborhood by RNA-mediated chromatin interactions (Fig. 3f). The models accurately illustrate the mechanisms pinpointed as causative for TAL1 overexpression based on extensive experimental testing [11]. This demonstrates that their inspection can give insights into the functional consequences of SVs.

The chromatin modeling method including SV information is provided as a web service at [28] together with a visualization tool.

Topological impact of structural variations

CTCF ChIP-seq data confirms that there are SVs which result in altered activity of reference interaction anchors. As an example, deletion chr14:35605439-35615196 of an interaction anchor leads to a significant depletion of CTCF signal in heterozygous samples and even to a complete vanishing of the signal in a homozygous sample (Fig. 4a). The CTCF signal drop reflects the lower or no potential of CTCF to bind to this segment. Therefore, in a cell line exhibiting the deletion, all of the chromatin contacts formed by this locus would not be present in one or both of the homologous chromosomes, depending on the genotype (Fig. 4b). The deletion is located in an intron of gene KIAA0391 but does not excise any coding sequence. Nevertheless, the genotypes show statistically significant differences in transcription rates of several genes (Fig. 4c). Even though the landscape of functional elements around the affected genes is complex to the extent that refrains from drawing definite conclusions, certain explanations may be proposed based on the changes of interaction patterns reflected in 3D models (Fig. 4d) and direct design of experiments. First, deletion chr14:35605439-35615196

removes a CTCF-mediated interaction anchor, which could be involved in the formation of insulated neighborhoods separating the PPP2R3C gene (upstream) from a group of enhancers (downstream). The loss of the putative insulated neighborhood boundary would promote higher activation of PPP2R3C (Fig. 4c), by allowing interactions between the gene and the enhancers. H3K4me1 signal, primarily associated with active enhancers [29], is notably stronger in deletion-affected homozygous GM18526 than in non-affected homozygous GM12878 in the enhancer region of interest (Fig. 4b and Additional file 2: Figure S7). This supports the proposed mechanism underpinning PPP2R3C increased activation. Moreover, the 3D distance between the PPP2R3C promoter and the strongest enhancer from the insulated neighborhood significantly decreases after introducing deletion in the 3D models (Fig. 4e) (see the “Methods” section). The existence of insulated neighborhoods is well established in the literature [11, 30, 31]. Second, deletion chr14:35605439-35615196 removes chromatin contact bringing the NFKBIA gene and one of the enhancers together in 3D space (Fig. 4d). This is reflected by the 3D distances between those two (Fig. 4e). The loss of the contact could explain lower NFKBIA expression in the samples carrying it (Fig. 4c). The association between NFKBIA transcription and genotype is not obvious as we did not find the difference in transcription between genotypes 0|0 and 1|1 statistically significant. However, we suspect that it would occur significant if the genotype 1|1 was represented by more samples than only 14. The deletion causes complex spatial rearrangements also around other genes, which contribute probably to the differences in their transcription rates between samples of different genotypes.

On the other hand, duplications of CTCF-mediated interacting genomic segments result in distinctively high relative values of CTCF signal in those segments in affected samples (Additional file 2: Figure S8A and S8B). The signal enrichment caused by those duplications supports a hypothesis that they create additional CTCF-binding loci with the potential to form additional long-range genomic contacts in the affected genomes.

Inversions in CTCF binding sites also modulate CTCF signal (Additional file 2: Figure S9A), which indicates they introduce changes in chromatin looping.

Apart from SVs disrupting the long-range chromatin interactions that join genomic segments located within one topological domain, there are examples of SVs modifying domain boundaries (Additional file 2: Figure S10A and S11A).

An aggregate analysis (see the “Methods” section) shows that CTCF ChIP-seq signals from samples having deletions (duplications) which intersect CTCF interacting anchors are depleted (enriched) in those anchors compared to signals from samples with the reference genotype (Additional file 2: Figure S12).

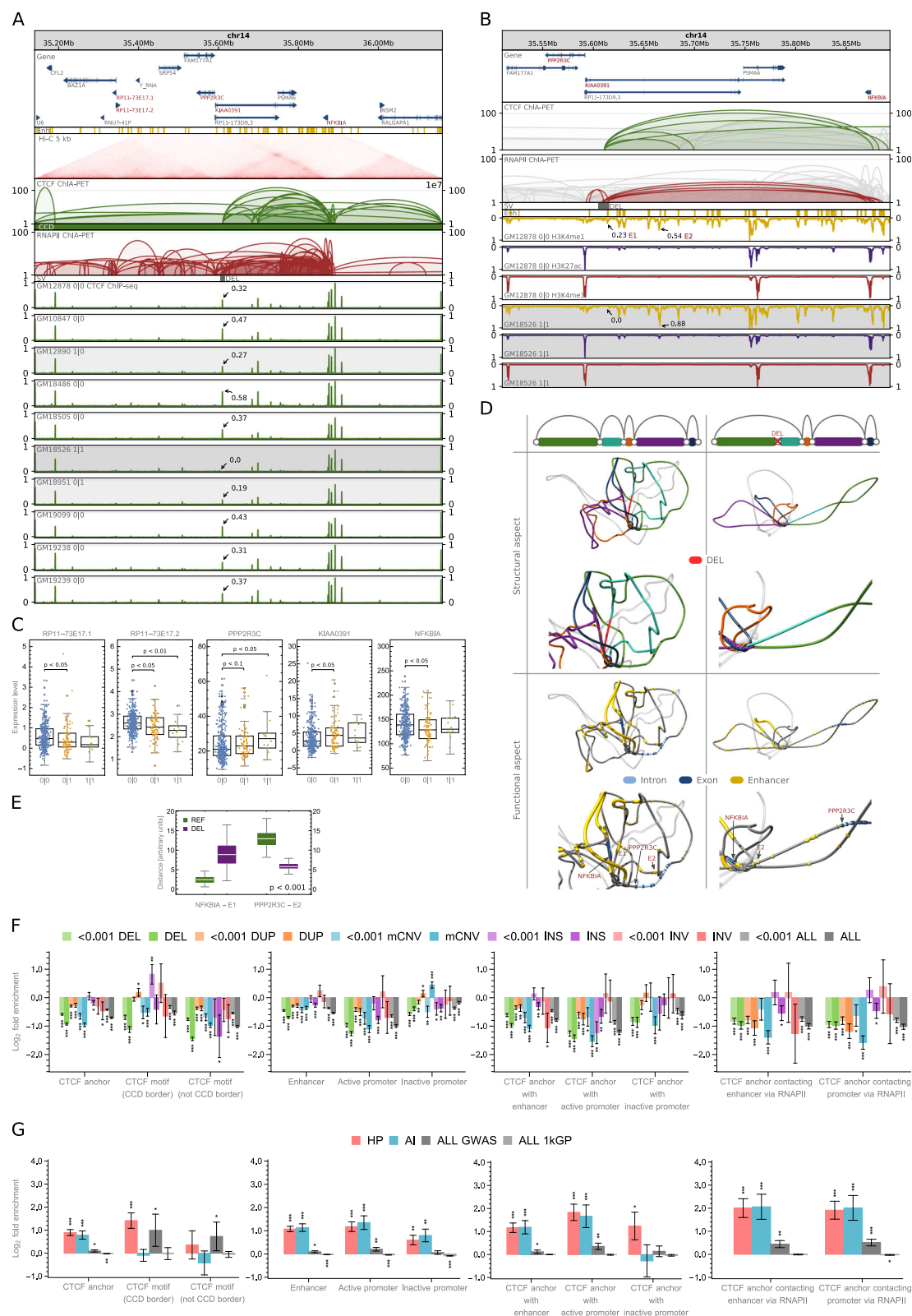


Fig. 4 (See legend on next page.)

(See figure on previous page.)

Fig. 4 Impact of SVs on genome organization at the population scale. **a** Browser view of a 1-Mb genomic segment with a deletion identified in a part of the human population. The deletion removes a CTCF anchor with enhancer located in an intron of KIAA0391. CTCF ChIP-seq signals from 10 lymphoblastoid cells of different genotypes are presented for comparison. For each track, ChIP-seq signal values (originally in RPMs) were divided by the maximal value of the signal in the visualized region. The highest signal peak in the genomic region covered by the deletion is marked in each signal track. **b** Close-up on ChIA-PET interactions at the deletion site displayed above the ChIP-seq profiles of histone modifications for GM12878—no deletion and GM18526—homozygous deletion. H3K4me1 is primarily associated with active enhancers, H3K27ac—with active promoters and enhancers, H3K4me3—with promoters. Compare with Additional file 2: Figure S7. **c** Differences in gene transcription rates between genotypes defined by the deletion. Genes exhibiting the differences in transcription which pass Mann-Whitney test with p value < 0.1 were reported, see Fig. 2c for box plot description. $n = 346, 85, 14$ sample points. **d** 3D models of the domain shown in **a** without the deletion (left column) and with the deletion (right column). Schematic drawings of loops shown in **b** (first row); 3D models with loops colored as on schematic drawings (second row); 3D models with NFKB1A and PPP2R3C genes (arrows are pointing toward the TSSs) and enhancers marked (third row). Every picture has its duplicated zooming in on the deletion site. **e** Distance in 3D Euclidean space between the NFKB1A promoter and enhancer E1 and between the PPP2R3C promoter and enhancer E2. In green, distribution of distances calculated in 3D models of the reference structure (REF), in purple—in models with the deletion introduced (DEL). For each case, 100 models were generated. The differences between REF and DEL groups are statistically significant (p values much less than 0.001), see Fig. 2c for box plot description. **f** Enrichment/depletion of genomic structural elements with SVs of different types and of different VAF (VAF < 0.001 and VAF ≥ 0.001). In case of CCD borders, only these fully imbedded in SV intervals are counted as affected, whereas for other structural elements ≥ 1 bp overlaps are counted. Error bars represent SD. **g** Enrichment/depletion of genomic structural elements with the 1000 Genomes Project SNPs (ALL 1kGP), all GWAS SNPs (ALL GWAS), GWAS SNPs associated with hematological parameters (HP), and with autoimmune diseases (AI). Error bars represent SD

Genotypes defined by such SVs exhibit significant differences in the expression of particular genes (Fig. 4c, Additional file 2: Figure S8C, S9B, S10D, and S11E).

We analyze the 3D structures of a part of those loci in Additional file 2: Figure S10B and S10C, Additional file 2: Figure S11C and S11D, and Additional file 2: Figure S13D.

To statistically assess the impact of SVs on the spatial organization of the genome, we analyzed their positions in relation to genomic structural elements like anchors of PET clusters, borders of CCDs, or gaps between them.

We observe that anchors of CTCF PET clusters are depleted of SVs (Fig. 4f) and that the rate of depletion is consistent among the loops of different directionality (Additional file 2: Figure S14).

We further identified CTCF-mediated interaction anchors intersected with enhancers and active and inactive gene promoters (see the “Methods” section). These anchors have a distinguished potential to play an important role in gene regulation. We observe that enhancers and promoters located in CTCF-mediated interaction anchors are significantly more conserved than the respective functional regions residing outside them (Fig. 4f). This indicates the importance of the genomic architecture mediated by CTCF in proper genome regulation. We additionally examined the conservation of CTCF anchoring sites, which interact with enhancers and gene promoters through RNAPII ChIA-PET contacts, and they also seem to be more conserved than the respective genomic functional elements (Fig. 4f).

Surprisingly, borders of CCDs do not seem to be distinctively well conserved. CTCF binding motifs we identified in CTCF ChIP-seq peaks outside CCD borders are significantly more depleted of SVs than the CTCF motifs indicating borders of CCDs (Fig. 4f). Moreover, CCD borders are enriched with rare insertions and are targets

of many duplications (Figs. 2e and 4f). There is also a slight enrichment of rare inversions in CCD borders, but because the set of inversions is small, the result is not statistically significant. However, we hypothesize that inverting the CTCF motifs at the borders of topological domains can be an important mechanism of genome reorganization and regulation. Six out of 786 inversions from the CHGV switch the directionality of CCD borders (Fig. 2e). Inversion chr10:15784798-15802449 is an example of such an event (Additional file 2: Figure S11A). It correlates with the transcription rate of a neighboring gene, VIM (Additional file 2: Figure S11E).

Stronger conservation of CTCF-mediated interaction anchors intersected with and connected to the known enhancers and promoters as compared to the conservation of enhancers and promoters located outside the anchors suggests that mutations of these anchors may lead to serious deregulations of gene transcription and can be related to a disease. To test this hypothesis, we intersected CTCF anchors with SNPs previously associated with disease in GWAS [32]. Having in mind the type of cell examined, we created separate sets of GWAS SNPs associated with hematological parameters and autoimmune diseases. Our analysis indeed shows a significant enrichment of these SNP classes in CTCF anchors intersected with enhancers and active promoters (Fig. 4g). Particularly, the enrichment is high in CTCF anchors being in RNAPII-mediated contact with enhancers and promoters. Both former and latter types of anchors are enriched with all GWAS SNPs. Importantly, enhancers and active promoters located outside the CTCF anchors are notably less enriched with GWAS SNPs than CTCF anchors associated with these functional elements (Fig. 4g). Generally, CTCF anchors and CCD boundaries are enriched with GWAS SNPs (Fig. 4g). Our observations

are consistent with the studies using capture Hi-C [33, 34] and additionally highlight the role of CTCF in shaping the network of functionally important genomic contacts.

We investigated particular examples of SNPs associated with autoimmune diseases (rheumatoid arthritis and vitiligo, rs4409785 T/C) and hematological parameters (red blood cell distribution width, rs57565032 G/T) (Additional file 2: Figure S15A and S16A). Both alter strongest CTCF binding motifs in the corresponding interaction anchors. However, their effect on CTCF binding is the opposite: rs4409785 increases the strength of CTCF motif it modifies (Additional file 2: Figure S15B), rs57565032 decreases (Additional file 2: Figure S16B). It is reflected in the CTCF signals corresponding to different genotypes (Additional file 2: Figure S15A and S16A). No other SNPs affect the CTCF motifs in those interaction anchors in presented genomes. Samples genotyped by these SNPs demonstrate significant differences in transcription rates of particular genes (Additional file 2: Figure S15C and S16C). One of them, MAML2, has been associated with cancer traits.

The already presented SNP rs12936231 (Fig. 2a) has been reported as a high-risk allele for asthma and autoimmune diseases and suggested to cause chromatin remodeling and alter transcription of certain genes, including ZBP2, GSDMB, and ORMDL3 [35]. We also found a correlation of genotypes set by rs12936231 with transcription rates of ZBP2, GSDMB, and ORMDL3 (Fig. 2c). Gene IKZF3, which also exhibits a correlation with rs12936231, has been related to B cell chronic lymphocytic leukemia.

Examples of SNPs in interacting anchors, but not associated with disease so far, can also be found. SNP rs60205880 alters CTCF-mediated chromatin looping and transcription of certain genes (Additional file 2: Figure S2). One of them, CCDC19, has been associated with bilirubin levels; another, IGSF8, is a member of an immunoglobulin superfamily. This demonstrates the potential of investigating genetic variants which target genomic structural elements for the identification of the mechanisms relating them to a disease.

The important question is how large the structural variation among healthy individuals is. Individual genomes sequenced in the 1000 Genome Project carry from 2571 to 6301 SVs, which affect from 1024 to 1419 CCDs and 55–347 CTCF anchors (Additional file 2: Figure S17). Almost all CCDs (98%) have an overlap with at least one SV from the CHGV. However, serious changes in local genome architecture are introduced by disruptions of interaction anchors rather than modifications of genomic regions between them. We identified 4944 unique patterns of SVs altering the interaction anchors in CCDs (we treat 2 patterns as identical if anchor-intersecting SVs they contain are the same; patterns are limited to single CCDs).

Together with the 2267 reference CCDs, it gives the number of CTCF-mediated topologies of genomic domains occurring in the 1000 Genomes Project population. We note that types of SVs are well separated in those patterns (Additional file 2: Figure S18). Eighty-seven percent of the patterns are comprised of only one SV type. There are 1539 patterns consisting of 2 or more SVs, and in 902 (59%) of them, all SVs are of the same type.

Population-specific topological alterations affected by structural variations

We additionally analyzed the intersections of SVs with genomic structural elements in the context of five continental groups: Africa (AFR), the Americas (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS). These populations are defined by the 1000 Genomes Project [36] (Additional file 1: Table S2).

In all the populations, deletions of interacting anchors are more frequent than duplications (Fig. 5a). This is not true for CCD borders (Fig. 5b), which agrees with the previous analyses showing that CCD borders are enriched with duplications (we note that there are significantly more deletions than duplications in the set of detected SVs (Fig. 2e)). Alterations of topological domain boundaries can be a general mechanism of genome structure evolution. The above results suggest that such a generic mechanism—similarly to the evolutionary process of introducing gene alterations by duplications—could use redundancy as a security measure. It could leave one chromatin loop with the original transcriptional function under evolutionary pressure, whereas the second could be acquiring novel local spatial landscape for genes and regulatory elements. This is in line with previous research on duplications [37, 38].

Our analysis shows that individual genomes from populations of African ancestry have the largest number of deletions in CTCF interaction sites (Fig. 5a). This is partly due to an outstanding number of all deletions identified in those genomes (Additional file 2: Figure S19). However, we still observe that African genomes, together with European genomes, have CTCF anchor sites less depleted of SVs unique to populations than genomes of other ancestries (Fig. 5c).

Interestingly, SVs found only in European genomes are significantly less depleted in CTCF interaction anchors intersected with enhancers or gene promoters than SVs unique to the rest of 5 distinguished continental groups (Additional file 2: Figure S20). We observe the same effect for borders of CCDs (Fig. 5c), but not for enhancers and promoters residing outside CTCF anchors (Additional file 2: Figure S21). This may suggest that part of the SVs identified in non-European populations overlap interaction sites specific for these populations and not observed in the reference 3D genome.

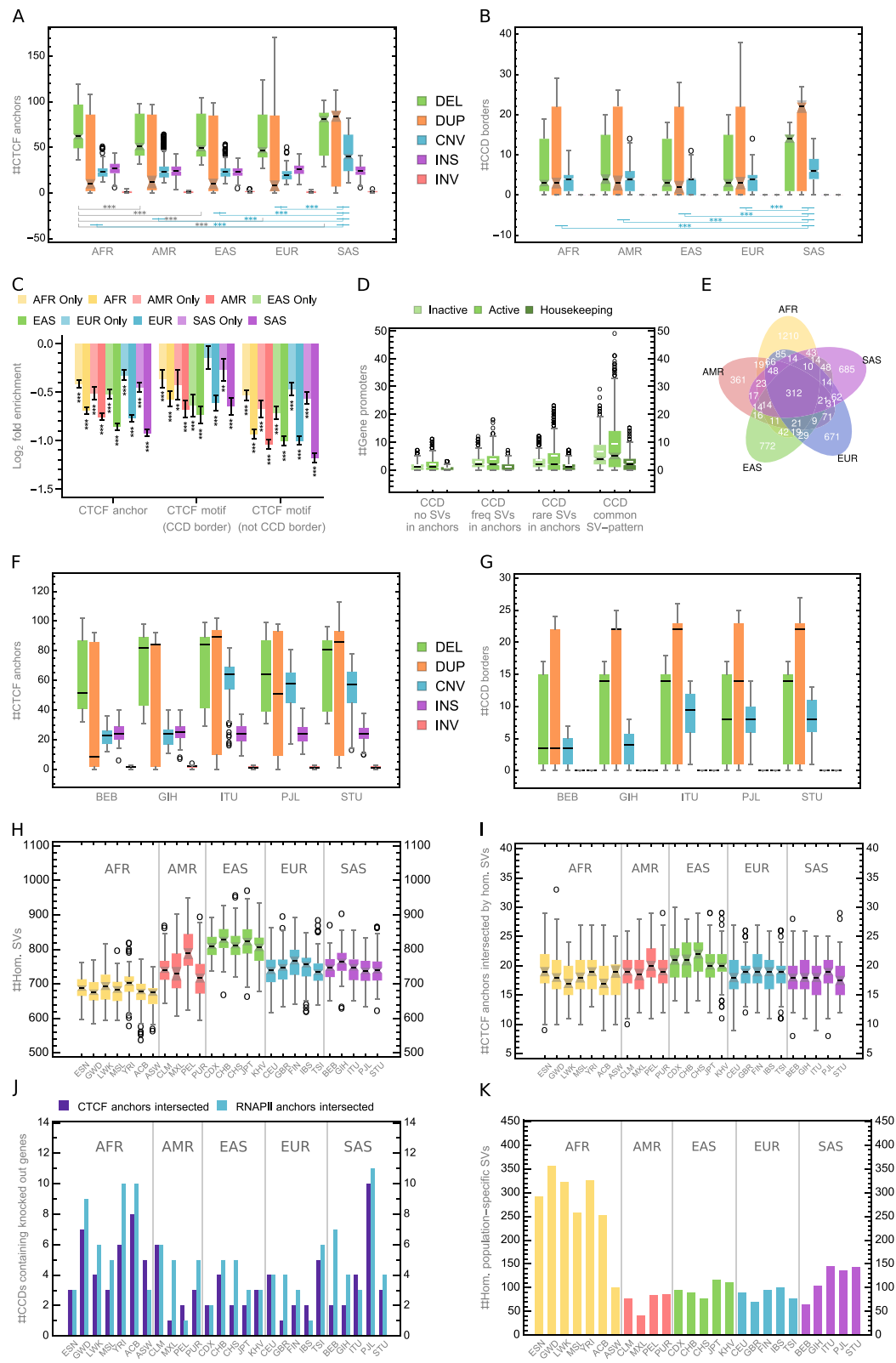


Fig. 5 (See legend on next page.)

(See figure on previous page.)

Fig. 5 Impact of population-specific structural variants on genome organization. **a** Number of CTCF anchors intersected by SVs of a given type identified in individuals from 5 continental groups, see Fig. 2c for box plot description. **b** Number of domain borders fully overlapped by SVs of a given type identified in individuals from 5 continental groups. **c** Enrichment/depletion of CTCF anchors and CCD boundaries with SVs divided by continental groups. CTCF motifs at CCD borders and outside CCD borders are shown for comparison. Only SVs fully covering motifs are counted as hits. **d** Number of gene promoters in domains covering regions in which SVs are identified. **e** CCD topology variability patterns by continental groups. **f** Number of CTCF anchors intersected by SVs of a given type identified in individuals from South Asian continental group. **g** Number of domain borders fully overlapped by SVs of a given type identified in individuals from South Asian continental group. **h** Number of homozygous SVs in individual human genomes by population. CNVs are treated as homozygous when the number of copies on both homologous chromosomes is different than in the reference (hom., homozygous). **i** Number of CTCF anchors intersected by homozygous SVs in individual genomes by population. **j** Number of CCDs containing human knockouts with CTCF (purple) or RNAPII (cyan) anchors intersected by homozygous population-specific SVs. **k** Homozygous SVs identified in a single human population

South Asian genomes, on the other hand, have a distinctively large number of duplicated CTCF anchor (Fig. 5a) and CCD border (Fig. 5b) sites. Whereas the distinctive number of altered structural elements in genomes of African ancestry could be expected based on the large genomic sequence variability in this population reported earlier [36], high structural variability in populations of South Asia is surprising. The ethnic groups which raise the statistics for South Asian continental group, especially those related to CNVs, are Indian Telugu in the UK (ITU), Punjabi in Lahore, Pakistan (PIL), and Sri Lankan Tamil in the UK (STU) (Fig. 5f, g). As a comparison, corresponding statistics for African and European continental groups seem to be more stable across the ethnic groups (Additional file 2: Figure S22). To investigate this further, we analyzed homozygous SVs. We hypothesized that the elevated number of structural changes observed in South Asian genomes could be caused by the high number of homozygous SVs that some of the populations in this continental group exhibit due to high consanguinity rates [39].

There are 13,767 homozygous SVs in the CHGV (we treat a CNV as homozygous, when there is a non-reference copy number on both homologous chromosomes). According to the data, genomes from East Asia, not South Asia, carry the largest number of the homozygous SVs (Fig. 5h). However, the differences in homozygous sequence variation are not reflected in the number of homozygously altered CTCF anchors. The latter seems not to be changing across populations (Fig. 5i).

The fruitful study of natural human knockouts performed on a cohort of 10,503 Pakistanis by the Human Knockout Project [40] made us investigate the homozygous SVs from the CHGV identified uniquely in a single population. We took the 1317 knocked out genes found in individuals from South Asia (in majority belonging to Urdu and Punjabi ethnic groups, over 70%) [40] and considered 656 CCDs they were located in. It turns out that homozygous SVs identified uniquely in Punjabi population intersect CTCF and RNAPII anchors in the largest number of CCDs (10 and 11 respectively) containing the

gene knockouts (Fig. 5j), even though a moderate number of population-specific homozygous SVs was found for this group (Fig. 5k). This suggests that gene knockouts may be accompanied (preceded, followed or assisted) by homozygous structural rearrangements.

For each of the continental groups, we prepared a list of patterns (similar to those described in the previous section) of anchor-intersecting SVs, which alter CCDs in the individuals belonging to this group. Even though most of the patterns are population-specific, we found 312 (6%) patterns common for all 5 continental groups (Fig. 5e). CCDs in which we found the common SV patterns are characterized by a particularly high number of gene promoters, including promoters of housekeeping genes (Fig. 5d). There are statistically more gene promoters in those CCDs than in other CCDs with modified anchors and in domains covering segments without changes in CTCF anchor sites. It is worth noticing that more promoters are located in CCDs containing CTCF anchors under variation than in those without them, which may suggest that the architecture of transcriptionally active genomic regions is more prone to mutation. CCDs with CTCF anchors under rare variation contain statistically more promoters of active genes than CCDs with CTCF anchors affected only by frequent SVs (Fig. 5d). Moreover, rare SVs happen to affect CTCF anchors in domains containing outstanding number of promoters (Fig. 5d). CCDs with rare variants in CTCF-interacting anchors can have up to 96 promoters of active genes (compared to 39 in CCDs with frequent SVs in anchors).

Regulation of gene transcription altered by topological variations in population

By combining information on chromatin interactions and population-scale genetic variation with transcriptome data from 462 lymphoblastoid cell lines gathered by the gEUVADIS Consortium [41, 42], we were able to draw first to our knowledge population-scale evidence-supported conclusions on the functional relation between SVs and genome architecture and provide a deeper insight into the functional role of genetic variation in the human genome.

The results of our analysis indicate that SVs influence gene transcription primarily by rearranging local looping structure of the genome.

For 445 out of the 462 samples provided by the gEUVADIS Consortium, there are also genotypes available in the 1000 Genomes Project database. We thus used PEER-normalized [41] gene expression levels of these 445 individuals for the association with their genotypes to identify expression quantitative trait loci (eQTLs). We use the term eQTL for any variation of genomic sequence which is identified as having an effect on the gene transcription level. The eQTL analysis was performed only with SVs, excluding SNPs.

We performed principal component analysis (PCA) on the expression data, which pinpointed 14,853 genes having the biggest variation in expression rates among individuals from all 23,722 genes present in the gEUVADIS dataset. We then related the expression levels of each of the 14,853 genes to genotypes (see the “Methods” section).

In the studies on eQTLs published so far [6, 41, 43–46], a genomic region of arbitrary size around a gene in question was conventionally set, and only the genetic variants located within this linear region were tested for being eQTLs for this gene. We argue that a more natural approach is to look for eQTLs within the whole topological domain the gene is located in. Therefore, for each of the selected genes, we evaluated the associations of its expression levels with all the genotyped SVs residing in the same CCD. For every gene-SV pair, least-square linear regression was performed and the significance of the slope was then tested in the permutation test. The resulting *p* values were adjusted for multiple testing to control the false discovery rate (FDR). We set a threshold of acceptance for $\text{FDR} \leq 10\%$ (see the “Methods” section).

We identified 234 unique SV-eQTLs modifying expression levels of 192 genes. The majority of the eQTLs found (55%) are deletions (Fig. 6a).

Earlier studies on eQTLs were limited in exploring the causal relation between genetic variation and gene expression to analyzing gene-variant and exon-variant intersections [43, 45], or the influence of genetic variation on transcription factor binding sites (TFBSs), transcription start sites (TSSs), or transcription end sites (TESs) [41, 46, 47]. In particular, one of the latest to our knowledge big study on the impact of structural variation on human gene expression reported that over 88% of predicted causal SVs did not alter gene structure or dosage [22]. The study showed enrichment of causal non-coding SVs in regions occupied by transcription factors or surrounding genes at distances up to 10 kb, but no deepened analysis of these regions was performed. Our analysis gives a broader idea of this relation and sheds light on the mechanisms through which SVs take part in genome regulation.

In agreement with [22, 46, 47], we observe an enrichment of eQTLs in TFBSs, but we see significantly higher enrichment of these in the genomic regions responsible for chromatin spatial organization. We divided the identified eQTLs into two sets: those located on the DNA chain closer than 17,800 bp to the genes they modify (proximal) and those located further apart (distal) (see the “Methods” section). The splitting value of 17,800 bp was chosen based on the distribution of RNAPII PET clusters lengths. It is a value for which the density of lengths of RNAPII clusters is equal to the half of the maximum density (Fig. 6b). The division is not exclusive—some of the eQTLs correlated with more than one gene are distal for one of them and proximal for other (Fig. 6c).

Active promoters and TFBSs are enriched with proximal eQTLs demonstrating their importance as gene-adjacent regulatory sites. Enhancers apart from being enriched with proximal eQTLs are enriched with the distal ones and represent regulatory elements interacting with genes by the nuclear space. However, the genomic elements most enriched with both proximal and distal eQTLs are anchors of RNAPII PET clusters (Fig. 6d). The abundance of eQTLs in the anchoring regions of the strong chromatin interactions mediated by RNAPII reaffirms the crucial role of this element of genome architecture in gene regulation.

As an example of eQTLs altering chromatin looping mediated by RNAPII, we investigate 5 deletions located in a HLA region (Fig. 6i). All of these deletions affect RNAPII interacting anchors (Fig. 6j) and are correlated with one or more of 5 HLA genes neighboring them (Fig. 6k). Three of the deletions are in a very strong linkage disequilibrium (LD) with each other (tested on the Central European population, Fig. 6h).

We hypothesize that proximal eQTLs modify TFBSs, TSSs, and TESs of genes as well as gene sequences but mostly they alter genes’ spatial contacts with regulatory elements and possibly with interaction centers, which has an immediate and straightforward influence on genes’ expression levels. Distal eQTLs have in turn higher potential to, apart from altering long-range RNAPII interactions, disrupt CTCF interactions that are longer than RNAPII-mediated chromatin loops (Fig. 6b) and shape the spatial structures of the whole topological domains.

Deletion chr1:248849861-248850138 is one of the eQTLs intersecting CTCF-mediated interaction anchors (Fig. 6a). Together with two other deletions (all in a very strong LD (Additional file 2: Figure S13F)), it introduces architectural changes correlated with increased transcription of a group of at least 6 olfactory receptor family genes, all residing on one chromatin loop (Additional file 2: Figure S13C). 3D models give more insight into the topological alterations induced by the deletions (Additional file 2: Figure S13D).

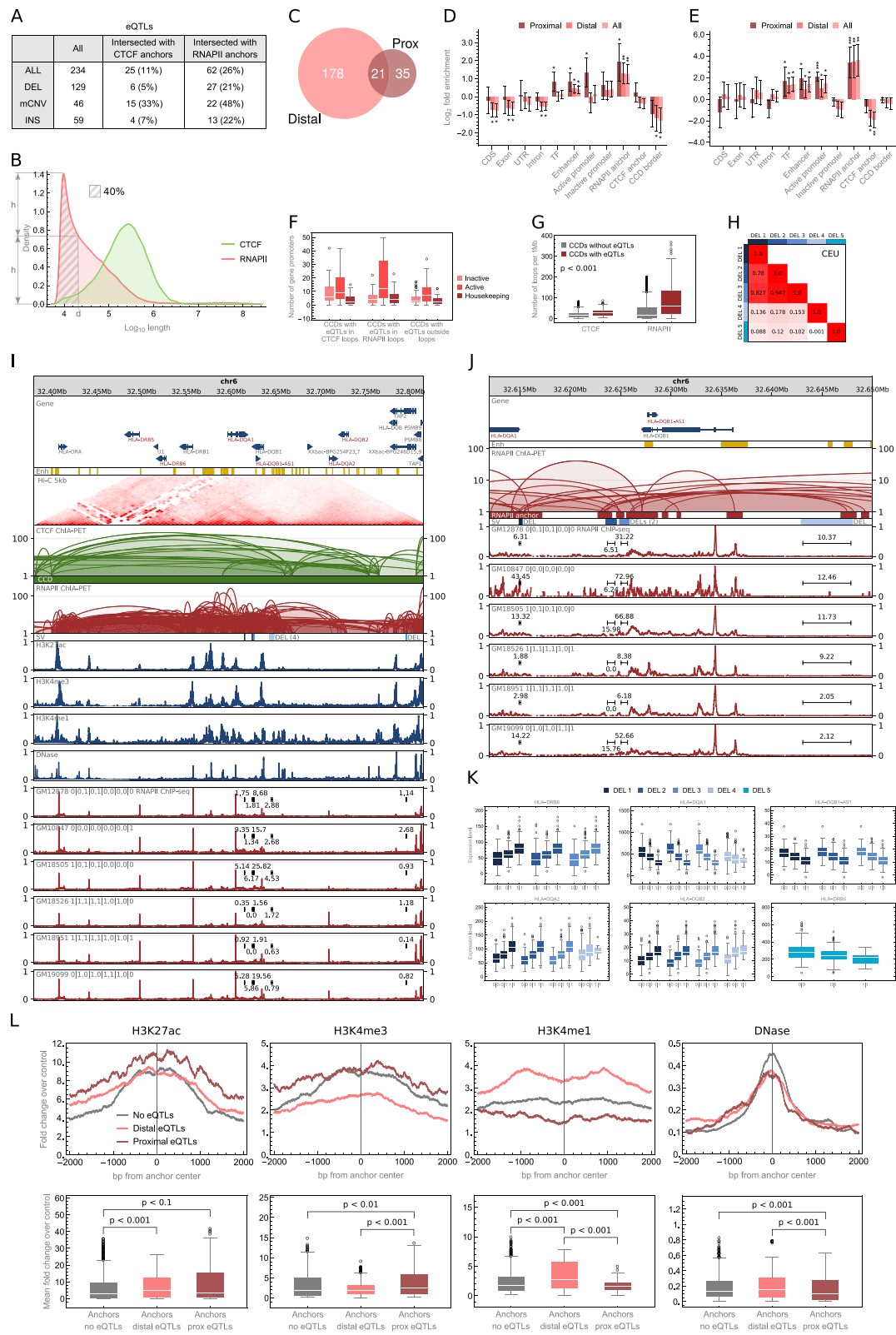


Fig. 6 (See legend on next page.)

(See figure on previous page.)

Fig. 6 Role of chromatin rearrangements in the regulation of gene transcription. **a** Table summarizing identified eQTLs and their intersections with interaction anchors. **b** Density plot showing genomic span distribution of PET clusters. d is the value (17,800 bp) by which eQTLs were split into proximal and distal. **c** Venn diagram showing the number of proximal (Prox) and distal eQTLs. **d** Enrichment/depletion of genomic elements with eQTLs. Error bars represent SD. **e** Enrichment/depletion of genomic elements with eQTLs of housekeeping genes. Error bars represent SD. **f** Abundance of gene promoters in CCDs, in which eQTLs were identified, see Fig. 2c for box plot description. $n = 16$ (CCDs with eQTLs in CTCF loops), 32 (CCDs with eQTLs in RNAPII loops), and 106 (CCDs with eQTLs outside loops) sample points. **g** Distributions of chromatin loop density in CCDs in which eQTLs were identified and in other CCDs. The density is measured for a particular CCD as an average number of CTCF-/RNAPII-mediated chromatin loops covering a 1-Mb fragment of this CCD. Differences between the groups are significant (p values < 0.001), see Fig. 2c for box plot description. $n = 2125$ (CCDs without eQTLs) and 142 (CCDs with eQTLs) sample points. **h** Linkage disequilibrium (measured as r^2 value in the CEU population) between deletions shown in **i** and **j**. Colors are assigned to the deletions as in **i-k**. **i** Browser view of a 0.4-Mb genomic segment with 5 deletions identified in a part of the human population, which disrupt RNAPII anchors and are eQTLs for 6 neighboring genes (signed with the red font). Each deletion has its color. RNAPII ChIP-seq signals from 6 lymphoblastoid cells of different genotypes are presented for comparison. For each track, normalized ChIP-seq signal values were divided by the maximal value of the signal in the visualized region. Sum of the signal values over the genomic regions occupied by the deletions is marked in each signal track. H3K27ac, H3K4me3, H3K4me1, and DNase-seq signal tracks from GM12878 are shown. **j** Close-up on the RNAPII-mediated interactions affected by 4 of the 5 deletions. Only the loops affected by the deletions are shown for clarity. **k** Genes which transcription is correlated with one or more of the deletions shown in **i** and **j** (p value < 0.001). Boxes with transcription rates associated with a particular deletion are marked with the color assigned to the deletion, as in **i** and **j**, see Fig. 2c for box plot description. $n = 132, 167, 146$ (DEL 1); 91, 208, 146 (DEL 2); 91, 208, 146 (DEL 3); 257, 158, 30 (DEL 4); 265, 154, 26 (DEL 5) sample points. **l** Signal strength of histone marks and DNase hypersensitivity sites in interaction anchors intersected with proximal eQTLs, distal eQTLs, and not intersected by eQTLs. For each mark, two plots are presented. A signal track around anchor center (± 2 kb) showing values for each genomic position averaged over all anchors from a given group (top). A box plot showing mean signal values in the same regions (bottom). Original signal values represent fold change over control. CTCF and RNAPII anchors were analyzed jointly, see Fig. 2c for box plot description. $n = 1000$ (anchors no eQTLs), 523 (anchors distal eQTLs), and 242 (anchors prox eQTLs) sample points

Another example of identified eQTL which alters CTCF-mediated chromatin structure is duplication chr17:44341412-44366497 (Additional file 2: Figure S23A). It duplicates the border of a CCD, and its emergence correlates with the transcription of the KANSL1-AS1 gene (Additional file 2: Figure S23B).

Even though we observed examples, we did not find anchors of CTCF PET clusters to be enriched with distal eQTLs (Fig. 6d). The fact we note, however, is that 17 of the identified eQTLs (24% of the anchor-intersecting eQTLs) intersect both RNAPII and CTCF anchors (Fig. 6a) and 36 (58%) of the eQTLs intersecting RNAPII anchors were detected in CCDs in which eQTLs targeting CTCF anchors were also found. This suggests that a change in gene expression observed among individuals can often be a result of a coordinated modification of RNAPII and CTCF anchors, but more investigation is needed to confirm this claim. Interestingly, CCDs in which eQTLs alter RNAPII anchors tend to embrace more active genes and housekeeping genes than CCDs with eQTLs not overlapping any interacting segments (Fig. 6f). On the other hand, CCDs with eQTLs in CTCF anchors contain many inactive genes (Fig. 6f).

Furthermore, we suspect that the enrichment analysis does not indicate that alterations of CTCF anchors significantly contribute to the variation of gene expression in population because the disruption of CTCF chromatin contacts would often provoke drastic changes in the local spatial organization of a genome not observed in healthy people. As we showed earlier, SNPs associated with disease favorably emerge in CTCF anchors (Fig. 4g).

For comparison with capture Hi-C (ChI-C) data, we mapped the identified eQTLs on genomic interactions reported in Mifsud et al. [33, 48]. Anchors of promoter-promoter and promoter-other ChI-C interactions were analyzed for the enrichment with the eQTLs (Additional file 2: Figure S24). The analysis shows that ChI-C anchors containing promoters are enriched with proximal eQTLs and depleted of the distal ones. A similar effect can be observed for ChIA-PET RNAPII anchors intersected with promoters (Additional file 2: Figure S24). However, unlike ChI-C anchors containing enhancers, ChIA-PET anchors intersected with enhancers are significantly enriched with distal eQTLs. The results for ChIA-PET data highlight the role of distal enhancers in gene regulation and may suggest that the interactions identified in RNAPII ChIA-PET are more transcriptionally active than the ones reported from ChI-C.

To state more firmly the relationship between proximal and distal eQTLs and chromatin activity, we collected (see the “Methods” section) sequencing (ChIP-seq) data for three histone modifications (H3K27ac, H3K4me3, H3K4me1) and information on chromatin accessibility (DNase-seq) and analyzed it in interaction anchors intersected with the eQTLs. H3K4me3 is primarily associated with promoters, H3K4me1 with active enhancers, and H3K27ac with active promoters and enhancers [29]. As expected, the interaction anchors altered by proximal eQTLs are enriched with promoter signal, whereas those affected by distal eQTLs with enhancer signal (Fig. 6l). This confirms that proximal eQTLs disrupt promoter-enhancer communication at the site of the promoter and distal eQTLs at the site of

the enhancer. The results are statistically significant, even though interaction anchors are enriched with chromatin marks in general (Fig. 6l) [17]. Furthermore, eQTLs emerge in densely connected genomic regions (Fig. 6g, l). This is also reflected by the fact that a single eQTL often intersects more than one RNAPII interaction anchor (Fig. 6j).

We repeated the eQTL analysis described above for housekeeping genes only (selected based on Eisenberg and Levanon [49]) to see if we find eQTLs for them and where the potential eQTLs are localized (see the “Methods” section). We found 36 unique eQTLs for 33 different housekeeping genes. None of the eQTLs is located within CTCF anchor, but we observe significant enrichment of them in RNAPII anchors (Fig. 6e). Therefore, there are differences in the expression rates of housekeeping genes among the samples, and they are mainly correlated with alternations of long-range chromatin contacts mediated by RNAPII.

On the other hand, we separately analyzed immune-related genes as genes specific to the lymphoblastoid cell lines (see the “Methods” section). Fourteen eQTLs were identified for these genes, out of which 4 intersect CTCF anchors. Three of these are anchors which contain enhancers and 1 contains a promoter region.

Two of the immune-related eQTLs (deletion chr22:39357694-39388574 and CNV chr22:39359355-39379392) cover the same CTCF anchor (Additional file 2: Figure S25A). Both are eQTLs for genes APOBEC3A, APOBEC3B, and CTA-150C2.16 (Additional file 2: Figure S25B), but the deletion completely excises APOBEC3B gene. In the presented samples (Additional file 2: Figure S25A), both of the SVs were identified, meaning that locus chr22:39357694-39388574 is (haplotype-specifically) excised in those genomes, which is reflected in CTCF signal for these samples.

Another example of an eQTL altering a CTCF anchor and regulating an immune-related gene is deletion chr17:73107713-73108273 (Additional file 2: Figure S26A). It is located over 750 kb apart from the correlated gene TRIM47 (Additional file 2: Figure S26B).

Whether cell type-specific genes are distinguished targets for SVs altering core chromatin architecture (as CTCF is believed to form the backbone network of genomic interactions) is an interesting question. However, more extensive testing has to be done to explore this hypothesis.

Discussion

It is already well established that part of the transcriptional variation between genomes or pathogenic phenotypes can be caused by chromatin topological alterations, but we still lack extensive genome-wide testing to assess the abundance and importance of these events. Our understanding of the importance of chromatin architecture in

genome regulation is based mainly on particular cases of SVs disrupting local 3D chromatin structure and subsequently leading to the deregulation of transcription of particular genes (in most of the studied cases associated with disease) [7, 9, 11]. There were more general insights into chromatin spatial rearrangements, but only in cancer genomes and at a less detailed level of the whole topological domains and their boundaries [8, 10]. No attempt was made so far to assess the abundance of chromatin architecture alterations in normal genomes, their functional impact, and the frequency with which genetic variations (related and not related with pathogenic phenotypes) target genomic regions responsible for the proper chromatin folding. The latter is specifically intriguing in the context of genome-wide association studies showing that over 95% of identified SNPs are located outside coding sequences [5]. This study is the first such attempt.

We mapped genetic variants identified in individuals from 26 human populations in the 1000 Genomes Project and disease-associated SNPs from GWAS onto chromatin three-dimensional structure of human lymphoblastoid cell line represented by CTCF and RNAPII ChIA-PET data. Our strategy for analyzing high-resolution CTCF and RNAPII genomic interaction data gives a comprehensive insight into the impact of SVs on chromatin organization. In agreement with previous studies [15, 50, 51], the analysis shows a high conservation of CTCF/cohesin-mediated chromatin topology between individual genomes. Moreover, the CTCF/cohesin-mediated promoter-enhancer interactions resulted as more conserved than enhancers and gene promoters not forming these type of interactions (Fig. 4f). This and the enrichment of disease-associated SNPs in CTCF/cohesin interaction anchors (Fig. 4g) indicate that they are critical mutational targets and that a significant part of non-coding regions of the genome targeted by disease-associated genetic variations is responsible for chromatin organization in cell nuclear space. On the other hand, alterations of chromatin interaction networks mediated by RNAPII are closely associated with the variation of gene transcription among population (Fig. 6d). The analysis of histone marks in RNAPII-mediated interaction segments targeted by transcription-associated SVs confirms that these SVs disrupt promoter-enhancer contacts (Fig. 6l). We found SVs correlated with gene transcription rates and disrupting local chromatin architecture built by RNAPII in a critical HLA region (Fig. 6i–k). We observe cases in which both CTCF and RNAPII anchors are modified in a topological domain containing genes with altered transcription rates, but the overall analysis indicates that the chromatin structure built by CTCF is strongly conserved across individuals and variation in gene transcription occurs by modifications of RNAPII interactions formed within this

structure. This hypothesis is consistent with the model presented by Tang et al. [17]. However, we suspect that the fact that we did not identify many examples of eQTLs located in CTCF anchors may also mean that the linear model used to detect eQTLs is too simple to account for complex nonlinear changes in gene transcription caused by modification of CTCF-mediated chromatin looping. This requires further investigation. Intriguingly, the evidence provided in this study identifies CTCF binding sites involved in the insulation of topological domains as frequently affected by duplications. This suggests that duplications of domain boundaries can have a distinguished role in evolutionary adaptation, similarly to duplications of genome coding sequences [37, 38].

We identified African and South Asian genomes as exhibiting the highest rates of structural variation in genomic interaction anchors and topological domain boundaries (Fig. 5a, b). South Asian genomes further stand out from the rest by having distinctively high numbers of mCNVs occurring in these genomic elements. This statistic is high within the continental group of South Asia mainly due to the input of three ethnic groups: Indian Telugu in the UK (ITU), Punjabi in Lahore, Pakistan (PJI) and Sri Lankan Tamil in the UK (STU) (Fig. 5f, g). We attempted to link the high rates of genome topology-affecting SVs observed in South Asia continental group to high consanguinity rates exhibited by some of the populations in this group. However, we did not detect an association between those two. A further investigation is needed including chromatin conformation capture experiments for different ethnic groups to address this question. On the other hand, we do note that SVs unique for the PJI ethnic group target genomic interaction anchors in the highest number of topological domains carrying knocked out genes found in the Human Knock-out Project from all the populations sequenced in the 1000 Genomes Project (Fig. 5j). This suggests that gene knockouts may be accompanied (preceded, followed, or assisted) by homozygous structural rearrangements. Interestingly, we observe that the rate of depletion of population-specific SVs in structural elements of the reference 3D genome is the smallest for European genomes (Fig. 5c and Additional file 2: Figure S20). Given that the 3D genome we use as the reference was obtained from a sample of European ancestry, it may suggest that part of the SVs specific for genomes of other ancestry target genomic interactions unique for those populations and not represented by the reference. This is an argument for more diversity in generating 3D genome data. Interestingly, domains which topology is affected by SV patterns occurring in genomes from all continental groups carry a distinctively high number of genes, including house-keeping genes (Fig. 5d). We generally observe that domains in which we identified SVs targeting CTCF/

cohesin-mediated interaction anchors carry significantly more genes than domains in which SVs occur only outside the anchors. These results may suggest that the replication process exhibits specificity in gene-rich genomic regions which causes common faults in copying sequences around genomic interaction sites which could be involved in this process. Another explanation could be that there is a set of topology-affecting SVs which occurred in gene-rich domains early in the evolution. The latter, however, does not explain the high abundance of genes in domains affected by rare SVs targeting CTCF/cohesin-mediated interaction anchors (Fig. 5d).

The analysis of spatially interacting genomic segments was possible using data from ChIA-PET experiments which identify such segments with high accuracy genome-wide. We applied additional filtering on CTCF ChIA-PET interacting segments, checking them for the co-occupancy by CTCF and cohesin ChIP-seq peaks, to obtain a highly credible set of CTCF-mediated chromatin interactions supported by cohesin. In case of CTCF ChIA-PET, it is even possible to identify individual CTCF binding motifs involved in the formation of the interactions, which brings high precision to the analysis. Such analysis would not be possible using Hi-C data in case of which genomic interaction segments are demarcated artificially, by segmenting the genome into adjacent bins of equal size and which resolution is rarely under the order of tens of kilobases, as a high resolution is obtained at the cost of very deep genome sequencing. The limitation of ChIA-PET experiment is that it captures chromatin interactions mediated by a particular protein, in contrast to Hi-C which identifies interactions of all kinds. Thus, Hi-C is a more complete representation of chromatin contacts present in the cell nucleus. However, CTCF ChIA-PET detects structural features exhibited by the non-specific Hi-C data. Data from these sources are highly correlated at the global whole-chromosome scale (Spearman's correlation coefficient in the range of 0.7–0.9) [17] and identify a very similar landscape of genomic structures at the local scale of topological domains (Fig. 1b) and chromatin loops (Fig. 1c). Moreover, the distinction between CTCF- and RNAPII-mediated interactions enabled us to spot the differences in the impact of SVs on them.

We used ChIA-PET data for GM12878 cell and treated it as a reference 3D genome for mapping SVs from other lymphoblastoid cell lines. CTCF and RNAPII ChIA-PET datasets for GM12878 are of the highest quality, and to our knowledge, no such datasets are currently available for any other lymphoblastoid cell line. Based on these highly credible genomic interactions and information on SVs identified in a population of 2504 human lymphoblastoid cells, we computationally predict chromatin interaction patterns for those cells. We claim that our computational tool is very useful for obtaining

individualized chromatin interaction patterns and in silico models of chromatin structures in the absence of experimental data, which generation requires expertise and certain money investments. On the other hand, because of the data unavailability, we could not confront our predictions with experimentally generated genomic interaction maps. No new biological experiments were conducted to support the correctness of our predictions, as this is purely computational analysis. However, we presented examples supported by available ChIP-seq datasets (Figs. 2a, 4a, and 6j; Additional file 2: Figure S2, S8, S9, S10, S11, S13, S15, S16, S23, S25, and S26) and genome-wide ChIP-seq analyses (Fig. 2d and Additional file 2: Figure S3 and S12) showing that predicting chromatin interactions based on ChIA-PET data for GM12878 and SVs from other lymphoblastoid cells is reasonable. Furthermore, we presented 3D models of an extensively studied (including the execution of CRISPR/Cas9 experiments) genomic region showing that their features are perfectly in line with discoveries and claims reported on this region [11] and that they could serve as accurate models for the mechanisms described in the earlier study. We do not claim that the models generated with our modeling method can alone explain the mechanisms underpinning associations of some SVs with gene transcription or constitute a proof of such mechanisms actually being the cause of observed changes in gene transcription rates. However, we believe that they can be a supporting tool in the analysis of potential disruptive effects of studied SVs on chromatin spatial organization and functional consequences of these alterations, helping to design a comprehensive study and to plan experiments more strategically.

We show that the topological variability of the human genome is rather limited (Additional file 2: Figure S18). However, because of the data used, the predictions are rather confined to lymphoblastoid cell lines. Nonetheless, our modeling method and web service providing the modeling tool can operate on uploaded data, if such data is at user's disposal.

Conclusions

This is the first genome-wide study on the influence of genetic variants on the chromatin organization and topological variability in the human population. It shows the critical impact of genetic variants on the higher-order organization of chromatin folding and provides a unique insight into the mechanisms regulating gene transcription at the population scale, among which the local arrangement of chromatin loops seems to be the leading one. This study highlights the importance and reason for further study on the role of chromatin architecture in genome regulation. It shows that further work on computational prediction of the chromatin 3D structures based on different factors changing among individuals is

required, as the emerging evidence shows that chromatin spatial organization is a crucial element to understand the genome regulation.

Methods

Genomic interactions

Genomic interactions analyzed in this study are 92,808 CTCF PET clusters and 100,263 RNAPII PET clusters identified by Tang et al. [17] for the GM12878 cell line (Additional file 1: Table S3 and S4) [18]. We refer the reader to this work for details on data processing pipeline used to find these interactions. Briefly, pair-end reads (PETs) sequenced in long-read ChIA-PET experiment were mapped to the human reference genome (hg19). Inter-ligation PETs were selected by the criterion of genomic span between the two ends of a PET exceeding 8 kb. Inter-ligation PETs overlapping at both ends were clustered together creating unique contacts (PET clusters) between 2 specific interaction loci, of strength equal to the size of the cluster. Anchors of CTCF PET clusters located within the distance of 500 bp along the DNA sequence were merged to more accurately correspond to loci covered by single CTCF binding peaks. This step led to further clustering of CTCF PET clusters and reduced their number from 92,808 to 80,157. Individual inter-ligation PET clusters and PET clusters of strength smaller than 4 are referred to as singletons.

ChIP-seq consensus peaks

We analyzed the directionality of CTCF interactions similarly to Tang et al. [17]. CTCF, SMC3 and RAD21 uniform ChIP-seq peaks available for the GM12878 cell line were downloaded from the ENCODE database (Additional file 1: Table S5) [52]. We extracted the 4-way consensus regions from all 4 sets of CTCF peaks to get highly credible CTCF-binding peaks. The same consensus was performed on SMC3 and RAD21 ChIP-seq segments to identify cohesin-binding peaks. Finally, 25,250 consensus regions from the CTCF and cohesin consensus peaks were obtained.

CTCF motif identification

We searched the CTCF/cohesin consensus peaks for CTCF-binding motifs. Nucleotide sequence of each of the CTCF/cohesin peaks was extracted from the hg19 assembly (downloaded from the UCSC database) using BEDTools (version 2.26.0) getfasta utility [53] and provided as an input to STORM (CREAD package version 0.84) [54]. Given the position weight matrix of a particular transcription factor-binding motif, STORM predicts the occurrences of the factor-binding motifs in provided DNA sequences. We performed the search with CTCF position weight matrix MA0139.1 downloaded from the JASPAR database [55] and found CTCF-binding motifs in

24,013 out of the 25,250 CTCF/cohesin consensus peaks. Only the motifs having a score higher than 0 were considered as valid, and for each peak, a motif with the highest score was selected.

Assigning orientation to CTCF loops

The CTCF motifs were overlapped with CTCF PET clusters. Forty-four thousand three hundred eighty out of the 80,157 CTCF PET clusters had both anchors overlapped by CTCF/cohesin consensus peaks (Additional file 1: Table S1), and only 2334 (3%) of them had no intersections with the consensus peaks at neither of sides. For 40,624 out of the 44,380 clusters (92%), at least one CTCF motif was found at both anchors. Thirty-three thousand sixty-two of these had exactly one motif at either side. PET clusters with anchors having more than one CTCF motif and of contradictory orientations were filtered out. From the 37,289 CTCF PET clusters with motifs of unique orientation in both anchors, 24,181 (65%) had motifs of convergent orientation at the two anchors, 6118 (16%) had motifs of tandem right orientation, 6089 (16%) PET clusters were of tandem left orientation, and 901 (2%) were of divergent orientation.

Chromatin contact domains

In this study, we used 2267 CTCF-mediated chromatin contact domains (CCDs) identified by Tang et al. (Additional file 1: Table S6) [18]. We refer the reader to this work for the details of CCD calling. Briefly, CCDs were identified by searching each chromosome for genomic segments continuously covered with CTCF PET clusters supported by CTCF/cohesin consensus peaks. Each identified CCD starts where the most upstream CTCF anchor from all the anchors of the CTCF PET clusters comprising the CCD starts and ends where the most downstream CTCF anchor ends. To define the borders of the CCDs more accurately, CTCF motifs found in the CTCF/cohesin consensus peaks and positioned within outermost anchors were identified. From these, the outermost CTCF motifs were selected as CCD borders. CTCF/cohesin consensus peaks with CTCF motifs were found in 4346 (96%) out of the 4534 outermost anchors. In case of the remaining 188 anchors, the strongest CTCF motifs identified in the full DNA sequence covered by the anchors were selected as indicators of CCD boundaries. Genomic regions complementary to CCDs (less hg19 reference genome assembly gaps) were defined as CCD gaps.

Enhancers and promoters

Definitions of enhancers used throughout this study were extracted from ChromHMM [56] hg19 annotations for the GM12878 cell line downloaded from the ENCODE database (Additional file 1: Table S5) [52]. Both weak and

strong enhancer annotations were adopted. Promoters were defined as ± 2 kb regions surrounding the gene transcription start sites (TSSs). The TSS coordinates were adopted from the GENCODE release 27 (mapped to hg19) [57, 58]. Only the promoters for protein-coding genes were considered. Promoters were defined as active if overlapped with anchors of RNAPII PET clusters or with RNAPII consensus ChIP-seq peaks and defined as inactive otherwise. RNAPII consensus peaks were obtained by performing consensus on 3 sets of RNAPII uniform ChIP-seq peaks available for the GM12878 cell line in the ENCODE database (Additional file 1: Table S5).

Enrichment analyses of SVs in genomic elements

In this study, we tested various genomic elements for the enrichment or depletion with structural variants (SVs). These tests were conducted according to a common scenario. Genomic elements of a given type represented by their positions in the hg19 reference genome were intersected with the positions of SVs. The ones having at least 1 bp overlap with at least 1 SV were counted. The genomic elements were then intersected with simulated SVs from 1000 sets generated by randomly shuffling positions of the original SVs, and the null distribution of counts of genomic elements overlapped with SVs was calculated. Each shuffled set contained the same number of elements in total and the same number of elements in subsets (deletions, duplications, etc.) as the real SV set. Elements of these sets were equally distributed on chromosomes and equally distributed in length to the real SVs. The operations of shuffling and intersecting genomic segments were performed with BEDTools (version 2.26.0) [53]. The enrichment or depletion of genomic elements overlapped with the real SVs compared to the same elements overlapped with randomly positioned simulated SVs was expressed as log₂ fold change of the number of the former versus the mean of the distribution of the number of the latter. Values of the measure were represented by the height of bars in the plots. Error bars in the plots show standard deviations of log₂ fold changes in each permutation test. To estimate the statistical significance of the test results, one-sided *p* values were calculated from the simulated distributions and marked above the bars by stars (3 stars, *p* value < 0.001; 2 stars, *p* value < 0.01; 1 star, *p* value < 0.1).

Subsets of SVs in enrichment analyses

For individual tests, the set of SVs was divided into various subsets. In particular, SVs with variant allele frequency (VAF) lower than 0.001 were considered separately in certain tests. In others, SVs were grouped according to the ancestry of individuals they were identified in, and sets of SVs emerging uniquely in one subpopulation were also created. The special set of SVs correlated with gene

expression (eQTLs) was subdivided into 2 sets: a set of eQTLs located closer on the DNA chain than 17,800 bp apart from the genes they modified and a set of eQTLs located further apart from their genes. The distances were calculated between TSSs (as defined in GENCODE version 12) [58] and centers of eQTL segments.

Subsets of GWAS SNPs

The set of GWAS SNPs used in this study was derived from the NHGRI-EBI GWAS Catalog, version from January 31, 2018 [32]. SNPs of traits associated with autoimmune diseases and hematological parameters were extracted as separate sets and mapped to dbSNP Build 150 for hg19 human genome assembly. SNPs mapping outside the main chromosome contigs, not having dbSNP ID or without coordinates on the hg19 and records containing multiple SNPs were excluded. This resulted in 2330 and 3919 unique SNPs associated with autoimmune diseases and hematological parameters respectively. For permutation tests with SNPs identified in healthy samples in the 1000 Genomes Project, we extracted a random sample of 1 million elements from the whole set of SNPs to limit the computation time and storage space.

Genomic elements in enrichment analyses

Analyzed in permutation tests, genomic elements associated with genes (annotated protein-coding sequence regions (CDSs), untranslated regions (UTRs) in protein-coding regions, exons, and introns) were adopted from the GENCODE release 27 (mapped to hg19). Permutation tests with eQTLs were an exception—in this case, gene elements from version 12 of GENCODE were used to maintain the consistency with the expression data which was analyzed with the earlier version of GENCODE. The positions of transcription factor-binding sites (TFBSs) were adopted from a file with uniform TFBS peaks downloaded from ENCODE (Additional file 1: Table S5).

Analysis of CTCF interaction anchors altered by SNPs

To test the impact of SNPs on the probability of CTCF binding to a CTCF anchor, we searched the nucleotide sequence of the anchor for CTCF motifs and compared the number and scores of these motifs with the CTCF motifs identified in the nucleotide sequence of this anchor after the introduction of alternative alleles. Only the motifs with a score higher than 0 were taken into consideration. Identification of CTCF motifs was performed as described in the “[CTCF motif identification](#)” section above.

We used ggseqlogo R package [59] to generate sequence logos from the frequency matrix MA0139.1 downloaded from the JASPAR database.

mRNA quantifications

PEER-normalized expression levels of 23,722 genes provided by the gEUVADIS Consortium [42] were used in our analyses. We refer the reader to Lappalainen et al. [41] for details on the process of transcriptome quantifications. In short, RNA-seq read counts over genes annotated in GENCODE (version 12) were calculated. This was done by summing all transcript RPKMs per gene. Read counts were corrected for variation in sequencing depth by normalizing to the median number of well-mapped reads among the samples and for technical noise. The latter was removed using PEER [60]. We logarithmized the corrected quantifications and standardized the distributions of transcription rates (for each gene individually).

Genotypes

Definitions of genomic sequence variations were taken from Sudmant et al. [25]. This SV set is a refined version of the callset released with the 1000 Genomes Project marker paper [36]. Only SVs (deletions, duplications, copy number variants, inversions, and insertions) were considered; SNPs were not included in the analysis. The genotype of an individual was represented as a sum of SV copies present on homologous chromosomes of the individual. Deletions were indicated by negative numbers. For example, if an individual had a deletion on both copies of a chromosome, the genotype was -2 . If it had 2 more copies of a genomic region (in relation to the reference genome) on one chromosome from the pair and 1 additional copy of this region on the second chromosome from the pair, the genotype was 3. Genotypes unchanged compared to the reference sequence (hg19 in this case) had codes 0. Genotypes of abundance lower than 1% in the studied population were neglected.

Linear models

The gEUVADIS Consortium provides RNA-seq data for 462 samples, but only a subset of these (445 samples) was genotyped in the 1000 Genomes Project. Thus, our analyses were performed on a population of 445 individuals for which both transcription and genotype data were available. Only the genotypes of abundance higher than 1% were considered. Sex chromosomes were excluded from the analyses. We took the logarithms of the PEER-normalized expression levels for calculations to correct it for far outliers and standardized the data. We started the analysis by performing principal component analysis (PCA) in the 23,722-dimensional space of gene expression rates. Based on the Scree Plot (Additional file 2: Figure S27), we decided to keep the first 100 principal components. Only the genes having contributions to these components not smaller than 0.01 were considered in the further analyses. By running this procedure, we got 14,853

genes of the largest contribution to the variance in gene transcription between samples. Every SV lying in the same CCD as one of these genes was tested for being eQTL for this gene. Least-squares linear regression between expression rates and genotypes was performed for each gene-SV pair. The slopes of the linear models were tested for statistical significance. First, for each linear model, two-sided p value was calculated in the test with a null hypothesis that slope is 0 (Wald test with t -distribution of the test statistics). Second, for each gene, we permuted the expression rates relative to genotypes 1000 times, recalculating at each iteration the linear regression for each gene-SV pair and recording the minimal p value among all pairs. Adjusted p values were calculated for each gene by dividing the ranks of the observed p values in the list of p values obtained in permutations by the number of permutations. Finally, to correct for multiple testing across genes, we applied the Benjamini-Hochberg procedure to the adjusted p values, estimating q values. At FDR 0.1, we found 192 genes with eQTLs. The same procedure was employed to identify eQTLs for housekeeping genes, except that PCA step was omitted. By mapping the names of housekeeping genes reported in Eisenberg and Levanon [49] on GENCODE (version 12), we obtained a list of 3784 genes. We found eQTLs for 33 of them. Lists of discovered eQTLs are provided in Additional file 1: Table S7 and S8.

Immune-related genes

Names and coordinates on the hg19 assembly of immunity genes were downloaded from InnateDB [61]. The gene names were mapped on GENCODE (version 12) for the purpose of the eQTL analysis. The final gene set contained 1051 elements.

ChIP-seq signal tracks

Raw sequencing data from ChIP-seq experiments published by Kasowski et al. [23] was processed to obtain signal tracks of CTCF and histone marks for 10 lymphoblastoid cell lines (GM12878, GM10847, GM12890, GM18486, GM18505, GM18526, GM18951, GM19099, GM19238, GM19239) [24]. The sequencing reads were aligned to the hg19 assembly using Bowtie2 (version 2.3.4.1) aligning tool [62]. The alignments were then passed to the bamCoverage utility from the deepTools2.0 (version 3.0.2) toolkit [63] to obtain RPM values genome-wide (the following command was evoked: `bamCoverage -b input.bam -o output.bw -of bigwig --binSize 10 --numberOfProcessors max/2 --normalizeUsing CPM --ignoreForNormalization chrX --extendReads --samFlagInclude 64`). Sequencing reads for every sample and for every experimental replicate were processed separately. Signals prepared for different biological replicates but for the same sample were merged to an averaged signal using the mean operator from the

WiggleTools1.2 package [64]. Each CTCF signal track included in the figures presents RPM values for a particular genomic region divided by the maximal value of the signal in this region.

The same post-alignment steps were applied to obtain signal tracks for SMC3 and RAD21 from the alignments downloaded from ENCODE (Additional file 1: Table S5).

H3K27ac, H3K4me3, H3K4me1, and DNase-seq data analyzed in the “Regulation of gene transcription altered by topological variations in population” section was downloaded from ENCODE in a form of bigWig files containing signal fold change over control (Additional file 1: Table S5).

Presented RNAPII signals were downloaded from the UCSC database (Additional file 1: Table S5).

Phased ChIP-seq signal tracks

Haplotype-specific CTCF and H3K4me1 ChIP-seq signals for 10 lymphoblastoid cell lines (GM12878, GM10847, GM12890, GM18486, GM18505, GM18526, GM18951, GM19099, GM19238, GM19239) were obtained from the raw sequencing data [24]. The reads sequenced for a particular cell line were aligned with Bowtie2 (version 2.3.4.1) aligning tool to the individualized nucleotide sequences of maternal and paternal chromosomes of this cell line. Only perfectly aligned reads were considered as valid. The sequences of maternal and paternal chromosomes were prepared with the vcf2diploid (version 0.2.6) tool from the AlleleSeq pipeline [65] using SNP phasing information from phase 3 of the 1000 Genomes Project. Additionally, the sequences of maternal and paternal chromosome 1 including SNP and SV phasing information from phase 3 of the 1000 Genomes Project were prepared. VCF files for chromosome 1 were processed by a custom script to represent alternative alleles as a sequence rather than SV identifier. Then, CTCF and H3K4me1 ChIP-seq data for 10 lymphoblastoid cell lines were mapped to those maternal and paternal sequences. To enable the comparison of phased ChIP-seq signals including SNP and SV information between individuals, aligned reads were remapped to hg19 reference with CrossMap (version 0.2.5) [66]. This step required chain files which were prepared as described in Minimal Steps For LiftOver [67]. Separate ChIP-seq signals for maternal and paternal chromosomes of the individual samples were calculated from the alignments prepared for the respective chromosomes analogously to the non-phased signals.

Linkage disequilibrium calculation

Linkage disequilibrium (LD) between selected SVs was calculated in the CEU population. Genotype information for 99 individuals from CEU population was extracted from phase 3 of the 1000 Genomes Project vcf files and passed to vcftools (version 0.1.15) [68] to convert it into

PLINK PED format (the following command was evoked: `vcftools -vcf input_sv.vcf --plink-tped --out plinksvs`). As some variants were multiallelic, the input vcf file was first processed with `bcftools` (version 0.1.19) to convert multiallelic variants to biallelic (the following command was evoked: `bcftools norm -m - -o sv.vcf -O v input_sv.vcf`). Then, LD between selected SVs measured as r^2 value was calculated with PLINK (version 1.07) [69] (the following commands were evoked: `plink --file plinksvs --make-bed --out out_sv; plink --bfile out_sv --r2 --ld-window-kb 1000 --ld-window 99999 --ld-window-r2 0.5`).

Modeling three-dimensional chromatin structures with 3D-GNOME

The ChIA-PET datasets typically consist of two types of interactions: high-frequency PET clusters (in the order of tens of thousands) representing strong, specific chromatin interactions, and singletons, numerous (in the order of tens of millions), but representing mostly non-specific and spurious contacts.

To make the best use of the information carried by these two distinct types of contacts, we employed a multiscale approach: first, we used the singletons to guide the low-resolution, megabase-scaled modeling, and then we used PET clusters to refine the obtained structures, achieving resolutions up to a few kilobases. We note that this approach is consistent with the widely accepted model of genome organization, in which the main roles are played by topological domains and chromatin loops. Here, at the stage of the low-resolution modeling, we attempt to position the topological domains relative to each other, and in the high-resolution, we model the position and shape of individual chromatin loops.

Low-resolution (chromosome level) modeling

The structure of a chromosome is represented using a “beads on a string” model. First, the chromosome is split into a number of approximately megabase-sized regions. Ideally, each region would correspond to a single topological domain. In practice, the split is made based on the patterns of PET clusters interactions (as a consequence, different regions typically will have different lengths (see [26] for details of the procedure)). Next, singleton heatmaps are created (much like the widely used Hi-C heatmaps, but with unequal bins). We treat an interaction frequency f_{ij} between a pair of regions i and j as a proxy of 3D distance d_{ij} between corresponding beads, assuming an inverse relationship $d_{ij} \sim c f_{ij}^{-\alpha}$, with α being a scaling exponent, and use Monte Carlo simulated annealing to position the beads to minimize the energy function

$E = \sum_{i,j} (d_{ij} - r_{ij})^2$, where r_{ij} is the actual distance between beads corresponding to regions i and j .

High-resolution modeling

In the second step, we model the shape and position of chromatin loops inside a single domain. We begin by splitting the interaction network given by PET clusters contained within a domain into a number of disjoint connected components that we call *blocks*. This allows us to model blocks independently. The modeling of each block is carried out in 2 steps. First, in the anchor step, we position the anchors of the loops identified by ChIA-PET relative to each other. The preferred distance between a pair of anchors i and j connected by a loop depends on the frequency of the PET cluster solely and is given by $d_{ij} = \delta + \alpha e^{-v(f_{ij}-\gamma)}$, where δ , α , v , and γ are all parameters (if the anchors are not connected, then d_{ij} is not specified). The energy function is identical in the form to the one used at the chromosome level, and we again use Monte Carlo simulations to find the optimal arrangements of the anchors. Then, in the subloop step, we keep the anchors' positions fixed, and we try to model the loops so that their shape, as well as relative position to other loops, best fit both the data and the physical constraints. Each loop is represented by k subanchor beads inserted between the neighboring anchors. We define stretching and bending energy terms as $E_s = \sum_i$

$(r_{i,i+1} - N_{i,i+1}^\beta)^2$ and $E_b = \frac{1}{2} \sum_i (1 - \hat{v}_{i-1,i} \cdot \hat{v}_{i,i+1})^2$, where

$N_{j,j+1}$ is a genomic distance between anchors j and $j+1$, $\hat{v}_{j,j+1}$ is a unit vector pointing from anchor j to $j+1$, and β is a constant parameter. To model the influence of short-range singleton interactions, we calculate the expected distances between all subanchor beads given only the physical constraints and then modify these distances based on the high-resolution singleton heatmaps for blocks. These updated distances are used in the third energy term $E_h = \sum_{i,j}$

$(d_{ij} - r_{ij})^2$, with d_{ij} and r_{ij} defined previously, but now for subanchor beads. The total energy function is simply $E_s + E_b + E_h$, and again, the Monte Carlo simulation is used for the optimization.

Modeling impact of SVs onto the three-dimensional chromatin structure

The algorithm modifies the reference genomic interactions and topological domains introducing information on SVs. The resulting genomic interaction data is then

passed to the 3D-GNOME modeling engine to obtain predicted 3D structures adjusted for SVs.

Anchors intersected by deletions are removed from the reference interaction pattern. As a result, all the interactions stemming from these anchors are eliminated yielding a structure with fewer loops, and with the loops directly neighboring the deletion being shorter or longer depending on the interaction pattern and the deletion size. If an outermost anchor in a CCD is intersected by a deletion, the boundary of the CCD is deleted and the CCD is fused with a neighboring CCD. CCDs covered by deletions are removed and the ones partially excised fused with neighboring CCDs. The anchors intersected by duplications are duplicated along with the contacts they have with other genomic segments in a way that interactions with anchors located upstream from the duplication are kept by the affected anchor and downstream interactions in respect to the duplication are acquired by the duplicate. The duplicate is positioned downstream from the affected anchor. If larger genomic fragments are duplicated, interactions between anchor duplicates are established equivalently to those between the duplicated anchors and anchor duplicates are not linked with the anchors of the original fragments. If a duplication expands over a CCD boundary, parts of the CCDs placed at the breakpoint after duplication are fused. Introducing duplications also results in elongation of loops overlapping the duplicated site. Inversions of genomic segments containing anchor sites result in changing positions of the anchors and its directionality relative to other anchors. After an anchor is inversed, we delete all its previous contacts and link it to the closest anchor with which it can form a convergent loop to reflect the preference of CTCFs to have symmetric conformation in dimers [15, 17]. In case of undirected protein targets, we link an inversed anchor with the closest anchor with no additional criteria. If the inversed anchor indicates a CTCF-mediated CCD border and it forms a convergent loop with the other border of the CCD, the border is removed. The anchors which lose all their connections as a consequence of SV introduction are linked with the closest anchor of orientation enabling the formation of convergent loop, in case of CTCF interaction networks, or closest anchor with no additional criteria in case of undirected protein targets. The insertions detected in the 1000 Genomes Project are almost solely insertions of transposable elements, which do not introduce new CTCF binding sites to the genome. Nevertheless, our algorithm enables introducing new CTCF binding sites to domain structures. Along with such insertions, new contacts are introduced between the inserted anchors and their neighbors.

The algorithm accounts also for SVs that miss the CTCF binding sites, but the introduction of these

results only in shortening or extending the corresponding chromatin loops.

Comparison of CTCF interaction segments among different individuals

In order to assess the fraction of CTCF interaction segments conserved among different lymphoblastoid cell lines, we calculated a number of CTCF anchors identified in GM12878, which were intersected with CTCF ChIP-seq peaks called in the remaining lymphoblastoid cell lines.

Since the available sets of CTCF ChIP-seq peaks for multiple lymphoblastoid cell lines [24] differed highly in size (Additional file 2: Figure S28), we performed an additional filtering on them. Only those CTCF ChIP-seq peaks which intersected with consensus CTCF binding sites were selected for each of the lymphoblastoid cell lines. The consensus CTCF binding sites were collected from the Ensembl Regulation 92 database [70]. They were identified by first performing a genome segmentation based on a variety of genome-wide assays from multiple cell types (including histone modification ChIP-seqs, TF ChIP-seqs, DNase-seq) and selecting segmentation state corresponding to CTCF peaks, and second, annotating the position of CTCF binding sites within the peaks using JASPAR position weight matrix MA0139.1, for more details, see the Ensembl website.

The consensus CTCF binding sites were downloaded via BioMart interface in genomic coordinates of hg38 assembly and converted to hg19 coordinates using UCSC liftOver tool [71].

Comparable datasets were obtained by the filtering (Additional file 2: Figure S29).

When using these filtered datasets, over 99% of interacting anchors occupied by CTCF peaks in GM12878 cell were identified as supported by CTCF peaks in each of the other lymphoblastoid samples. However, we note that a similar rate of 83% and higher is observed when using unfiltered sets of CTCF ChIP-seq peaks (Additional file 2: Figure S30).

Aggregate analysis of ChIP-seq signals in altered interaction anchors

In order to analyze the overall behavior of ChIP-seq signal in interacting genomic segments affected by deletions or duplications, we performed an aggregate analysis. The same procedure was applied regardless of SV type (deletion or duplication) and ChIP-seq target protein (CTCF or RNAPII). We describe it using an example of CTCF interacting segments affected by deletions.

First, CTCF anchors intersected by deletions exhibited by at least one of the 10 lymphoblastoid cell lines with available CTCF ChIP-seq data [24] were identified. For each such anchor, 200 bins were defined—100 bins of

equal size covering an anchor and 50 equal-size bins covering genomic regions 500 bp upstream and downstream from the anchor. Averaged raw CTCF ChIP-seq signal was calculated in every bin for every sample. For each sample, a mean of the signal over the whole genome was found and subtracted from the extracted binned signal values. The maximal mean from all the samples was then added to the values to make them positive. Obtained values were then averaged over the samples exhibiting and not exhibiting the deletion. The log₂ of the ratio of the signal values obtained for the first group to the ones obtained for the second group was calculated. The log₂ fold changes calculated for all the anchors affected by deletions were then averaged and plotted in Additional file 2: Figure S12A.

Additional files

Additional file 1: Supplemental tables. (XLSX 12146 kb)

Additional file 2: Supplemental figures. (PDF 17323 kb)

Acknowledgements

Not applicable

Authors' contributions

DP and MS conceived and implemented the methodology of modeling CCDs of individual genomes at the population scale. PS, DP, ZT, and YR devised the 3D-GNOME used as the main engine for modeling. MW extended the 3D-GNOME web service to provide the SV-including modeling method (3D-GNOME 2.0). MS designed the statistical analysis part. MS and AK performed the analyses. MS, AK, and ZT extracted and prepared the data for the analyses. MS, PS, AK, ZT, YR, and DP prepared the manuscript. MS was a major contributor in writing the manuscript. DP, YR, and ZT supervised the study. All authors read and approved the final manuscript.

Funding

This work was carried out within the TEAM program of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund, co-supported by Polish National Science Centre (2014/15/B/ST6/05082), and the grant 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation" within 4D Nucleome NIH program.

Availability of data and materials

The authors declare that the data supporting the findings of this study are available within the supplemental files or in public repositories to which references are given in the paper and supplemental information files. ChIA-PET dataset supporting the conclusions of this article was downloaded from the Gene Expression Omnibus (GEO), accession number GSE72816, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72816> [18]. Positions of the CTCF- and RNAII-mediated chromatin interactions and chromatin contact domains in the hg19 reference genome assembly are additionally provided in Additional file 1: Table S3, S4, and S6 respectively. Hi-C data analyzed in this study was downloaded from GEO, accession number GSE63525, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525> [21]. CTCF, SMC3, RAD21 and RNAII ChIP-seq datasets used for chromatin interactions filtering and support were downloaded from The Encyclopedia of DNA Elements (ENCODE) [52]; the accession numbers and URLs of these datasets are provided in Additional file 1: Table S5. The set of ChIA-PET interactions filtered by the co-occupancy by CTCF and cohesin (SMC3 and RAD21 subunits) is provided in Additional file 1: Table S1. CTCF and histone marks in ChIP-seq datasets for multiple lymphoblastoid cell lines were downloaded from GEO, accession number GSE50893, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50893> [24]. RNAII ChIP-seq signals

for multiple lymphoblastoid cell lines were downloaded from ENCODE; detailed accession information is provided in Additional file 1: Table S5. Structural variants analyzed in this study were adopted from phase 3 of the 1000 Genomes Project, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz [25]. NHGRI-EBI GWAS Catalog released on January 31, 2018, was analyzed in this study, <http://ftp.ebi.ac.uk/pub/databases/gwas/releases/2018/01/31/gwas-catalog-associations.tsv> [32]. RNA-seq dataset used for eQTL analysis was downloaded from ArrayExpress, accession number E-GEUV-1, <https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/> [42]. Genomic interactions identified in the GM12878 cell line through Capture Hi-C experiments were used for comparison. The dataset was downloaded from ArrayExpress, accession number E-MTAB-2323, <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2323/> [48]. Gene annotations from GENCODE versions 12 and 27 were used throughout the study [58]. Immune-specific genes were selected based on the InnateDB annotation, <https://www.innate-db.com/annotatedGenes.do?type=innate-db> [61]. Chromatin state segmentation for the GM12878 cell line by ChromHMM and TFBS clusters were downloaded from ENCODE and the detailed accession information is provided in Additional file 1: Table S5.

Python code developed to include SV information in chromatin interaction patterns is available at https://bitbucket.org/4dnucleome/spatial_chromatin_architecture under the BSD 2-Clause License. The version used in the manuscript is deposited in zenodo <https://doi.org/10.5281/zenodo.2837248> [72]. The algorithm was integrated with the 3D-GNOME modeling engine [26] and a visualization tool into a web service (3D-GNOME 2.0) [28].

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

¹Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland. ²Faculty of Physics, University of Warsaw, Pasteura 5, 02-093 Warsaw, Poland. ³Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland. ⁴Centre for Innovative Research, Medical University of Białystok, Kilinskiego 1, 15-089 Białystok, Poland. ⁵l-BioStat, Hasselt University, Agoralaan building D, BE3590 Diepenbeek, Belgium. ⁶Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China. ⁷The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, USA.

Received: 24 October 2018 Accepted: 30 May 2019

References

- Malhotra D, Sebat J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*. 2012;148:1223–41.
- Stankiewicz P, Lupski J. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010;61:437–55.
- Zollino M, Orteschi D, Murdolo M, Lattante S, Battaglia D, Stefanini C, Mercuri E, Chiurazzi P, Neri G, Marangi G. Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. *Nat Genet*. 2012;44:636–8.
- Talkowski M, Mullegama S, Rosenfeld J, van Bon W, Shen Y, Repnikova J, Gastier-Foster J, Thrush D, Kathiresan S, Ruderfer D, et al. Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. *Am J Hum Genet*. 2011;89:551–63.
- Maurano M, Humbert R, Rynes E, Thurman R, Haugen E, Wang H, Reynolds A, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337:1190–5.
- Sudmant P, Rausch T, Gardner E, Handsaker R, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz M, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526:75.
- Lupianez D, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Hom D, Kayserili H, Opitz J, Laxova R, et al. Disruptions of topological chromatin

- domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. 2015;161:1012–25.
8. Weischenfeldt J, Dubash T, Drains A, Mardin B, Chen Y, Stutz A, Waszak S, Bosco G, Halvorsen A, Raeder B, et al. Pan-cancer analysis of somatic copy-number alterations implicates *IRS4* and *IGF2* in enhancer hijacking. *Nat Genet*. 2017;49:65–74.
 9. Northcott P, Buchhalter J, Morrissy A, Hovestadt V, Weischenfeldt J, Ehrenberger T, Grobner S, Segura-Wang M, Zichner T, Rudneva V, et al. The whole-genome landscape of medulloblastoma subtypes. *Nature*. 2017;547:311–+.
 10. Dixon J, Xu J, Dileep V, Zhan Y, Songs F, Le V, Yardimci G, Chakraborty A, Bann D, Wang Y, et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet*. 2018;50:1388.
 11. Hnisz D, Weintraub A, Day D, Valtou A, Bak R, Li C, Goldmann J, Lajoie B, Fan Z, Sigova A, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*. 2016;351:1454–8.
 12. Bianco S, Lupianez D, Chiariello A, Annunziata C, Kraft K, Schopflin R, Wittler L, Andrey G, Vingron M, Pombor A, et al. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat Genet*. 2018;50:662.
 13. Spielmann M, Lupianez D, Mundlos S. Structural variation in the 3D genome. *Nat Rev Genet*. 2018;19:453–67.
 14. Lieberman-Aiden E, van Berkum N, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie B, Sabo P, Dorschner M, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326:289–93.
 15. Rao S, Huntley M, Durand N, Stamenova E, Bochkov I, Robinson J, Sanborn A, Machol I, Omer A, Lander E, Aiden E. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159:1665–80.
 16. Fullwood M, Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem*. 2009;107:30–9.
 17. Tang Z, Luo O, Li X, Zheng M, Zhu J, Szalaj P, Trzaskoma P, Magalska A, Włodarczyk J, Ruszczycki B, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*. 2015;163:1611–27.
 18. Luo O, Tang Z, Li X, Ruan Y. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Gene expression omnibus*. 2015. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72816>. Accessed 17 Jan 2019.
 19. Ong C, Corces V. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet*. 2014;15:234–46.
 20. Ou H, Phan S, Deerinc T, Thor A, Ellisman M, O'Shea C. ChromEMT: visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*. 2017;357. <https://science.sciencemag.org/node/697147.full>.
 21. Rao S, Huntley M, Lieberman Aiden E. A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping. *Gene Expression Omnibus*. 2014. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525>. Accessed 10 Sept 2018.
 22. Chiang C, Scott A, Davis J, Tsang E, Li X, Kim Y, Hadzic T, Damani F, Ganel L, Montgomery S, et al. The impact of structural variation on human gene expression. *Nat Genet*. 2017;49:692.
 23. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg J, Kundaje A, Liu Y, Boyle A, Zhang Q, Zakharia F, Spacek D, et al. Extensive variation in chromatin states across humans. *Science*. 2013;342:750–2.
 24. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg J, Kundaje A, Liu Y, Boyle A, Zhang Q, Zakharia F, Spacek D, et al. Extensive variation in chromatin states across humans. *Gene Expression Omnibus*. 2013. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50893>. Accessed 3 Mar 2019.
 25. Sudmant P, Rausch T, Gardner E, Handsaker R, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz M, et al. An integrated map of structural variation in 2,504 human genomes. IGS: the international genome sample resource. 2015. <http://www.1000genomes.org/phase-3-structural-variant-dataset>.
 26. Szalaj P, Tang Z, Michalski P, Pietal M, Luo O, Sadowski M, Li X, Radew K, Ruan Y, Plewczynski D. An integrated 3-Dimensional Genome Modeling Engine for data-driven simulation of spatial genome organization. *Genome Res*. 2016;26:1697–709.
 27. de Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*. 2013;502:499–506.
 28. 3D-GNOME 2.0 - 3D chromatin organization web service. <https://3dgenome.cent.uw.edu.pl>. Accessed 15 May 2019.
 29. Ernst J, Kheradpour P, Mikkelsen T, Shores N, Ward L, Epstein C, Zhang X, Wang L, Issner R, Coyne M, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011;473:43–U52.
 30. Downen J, Fan Z, Hnisz D, Ren G, Abraham B, Zhang L, Weintraub A, Schuijers J, Lee T, Zhao K, Young R. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*. 2014;159:374–87.
 31. Phillips-Cremis J, Sauria M, Sanyal A, Gerasimova T, Lajoie B, Bell J, Ong C, Hookway T, Guo C, Sun Y, et al. Architectural protein subclasses shape 3D Organization of Genomes during lineage commitment. *Cell*. 2013;153:1281–95.
 32. Buniello A, MacArthur J, Cerezo M, Harris L, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Solis E, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies (release 2018/01/31). NHGRI-EBI GWAS Catalog. 2019. <ftp://ftp.ebi.ac.uk/pub/databases/gwas/releases/2018/01/31/gwas-catalog-associations.tsv>.
 33. Mifsud B, Tavares-Cadete F, Young A, Sugar R, Schoenfelder S, Ferreira L, Wingett S, Andrews S, Grey W, Ewels P, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*. 2015;47:598–606.
 34. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, Cooper N, Barton A, Wallace C, Fraser P, et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat Commun*. 2015;6:10069.
 35. Verlaan D, Berlivet S, Hunninghake G, Madore A, Larivière M, Moussette S, Grundberg E, Kwan T, Ouimet M, Ge B, et al. Allele-specific chromatin remodeling in the ZBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. *Am J Hum Genet*. 2009;85:377–93.
 36. Altshuler D, Durbin R, Abecasis G, Bentley D, Chakravarti A, Clark A, Donnelly P, Eichler E, Flicek P, Gabriel S, et al. A global reference for human genetic variation. *Nature*. 2015;526:68.
 37. Dennis M, Eichler E. Human adaptation and evolution by segmental duplication. *Curr Opin Genet Dev*. 2016;41:44–52.
 38. Dennis M, Harshman L, Nelson B, Penn O, Cantalieri S, Huddleston J, Antonacci F, Penewit K, Denman L, Raja A, et al. The evolution and population diversity of human-specific segmental duplications. *Nat Ecol Evol*. 2017;1(3):69.
 39. Bittles A, Mason W, Greene J, Rao N. Reproductive-behavior and health in consanguineous marriages. *Science*. 1991;252:789–94.
 40. Saleheen D, Natarajan P, Armean I, Zhao W, Rasheed A, Khetarpal S, Won H, Karczewski K, O'Donnell-Luria A, Samocha K, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*. 2017;544:235.
 41. Lappalainen T, Sammeth M, Friedlander M, 't Hoen P, Monlong J, Rivas M, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira P, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501:506–11.
 42. Lappalainen T, Sammeth M, Friedlander M, 't Hoen P, Monlong J, Rivas M, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira P, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *ArrayExpress*. 2013. <https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/>. Accessed 14 May 2018.
 43. Schlattl A, Anders S, Waszak S, Huber W, Korbel J. Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res*. 2011;21:2004–13.
 44. Stranger B, Forrest M, Dunning M, Ingle C, Beazley C, Thorne N, Redon R, Bird C, de Grassi A, Lee C, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007;315:848–53.
 45. Pickrell J, Marioni J, Pai A, Degner J, Engelhardt B, Nkadori E, Veyrieras J, Stephens M, Gilad Y, Pritchard J. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010;464:768–72.
 46. Veyrieras J, Kudravalli S, Kim S, Dermizakis E, Gilad Y, Stephens M, Pritchard J. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*. 2008;4(10):e1000214.
 47. Gaffney D, Veyrieras J, Degner J, Pique-Regi R, Pai A, Crawford G, Stephens M, Gilad Y, Pritchard J. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol*. 2012;13(1):R7.
 48. Mifsud B, Tavares-Cadete F, Young A, Sugar R, Schoenfelder S, Ferreira L, Wingett S, Andrews S, Grey W, Ewels P, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *ArrayExpress*. 2015. <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2323/>. Accessed 11 May 2018.
 49. Eisenberg E, Levanon E. Human housekeeping genes, revisited (vol 29, pg 569, 2013). *Trends Genet*. 2014;30:119–+.

50. Dixon J, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu J, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485:376–80.
51. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science*. 2002;295:1306–11.
52. Dunham I, Kundaje A, Aldred S, Collins P, Davis C, Doyle F, Epstein C, Fietze S, Harrow J, Kaul R, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
53. Quinlan A, Hall I. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
54. Schones D, Smith A, Zhang M. Statistical significance of cis-regulatory modules. *BMC Bioinformatics*. 2007;8:19.
55. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon J, van der Lee R, Bessy A, Cheneby J, Kulkarni S, Tan G, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res*. 2018;46:D260–6.
56. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9:215–6.
57. Frankish A, Diekhans M, Ferreira A, Johnson R, Jungreis I, Loveland J, Mudge J, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47:D766–73.
58. Frankish A, Diekhans M, Ferreira A, Johnson R, Jungreis I, Loveland J, Mudge J, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. GENCODE. 2019. ftp://ftp.ebi.ac.uk/pub/databases/genocode/Gencode_human/release_12/genocode.v12.annotation.gtf.gz, ftp://ftp.ebi.ac.uk/pub/databases/genocode/Gencode_human/release_27/GRCh37_mapping/genocode.v27lift37.annotation.gtf.gz. Accessed 14 May 2018.
59. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*. 2017;33:3645–7.
60. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*. 2012;7:500–7.
61. Breuer K, Foroushani A, Laird M, Chen C, Sribnaia A, Lo R, Winsor G, Hancock R, Brinkman F, Lynn D. InnateDB: systems biology of innate immunity and beyond - recent updates and continuing curation (downloaded 04/2018). InnateDB. 2013. <https://www.innatedb.com/annotatedGenes.do?type=innatedb>. Accessed 16 Apr 2018.
62. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–U354.
63. Ramirez F, Ryan D, Gruning B, Bhardwaj V, Kilpert F, Richter A, Heyne S, Dunder F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 2016;44:W160–5.
64. Zerbino D, Johnson N, Juettemann T, Wilder S, Flicek P. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics*. 2014;30:1008–9.
65. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol*. 2011;7:522.
66. Zhao H, Sun Z, Wang J, Huang H, Kocher J, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*. 2014;30:1006–7.
67. Minimal Steps For LiftOver. http://genomewiki.ucsc.edu/index.php/Minimal_Steps_For_LiftOver. Accessed 10 Mar 2019.
68. Danecek P, Auton A, Abecasis G, Albers C, Banks E, DePristo M, Handsaker R, Lunter G, Marth G, Sherry S, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–8.
69. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly M, Sham P. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75.
70. Zerbino D, Achuthan P, Akanni W, Amode M, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron C, et al. Ensembl 2018. *Nucleic Acids Res*. 2018;46:D754–61.
71. Hinrichs A, Karolchik D, Baertsch R, Barber G, Bejerano G, Clawson H, Diekhans M, Furey T, Harte R, Hsu F, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*. 2006;34:D590–8.
72. Sadowski M. Spatial chromatin architecture alteration by structural variations in human genomes at the population scale [code]. 2019. <https://doi.org/10.5281/zenodo.2837248>. Accessed 15 May 2019.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome

Michał Wlasnowolski^{1,2}, Michał Sadowski¹, Tymon Czarnota², Karolina Jodkowska¹, Przemysław Szalaj^{1,3,4}, Zhonghui Tang⁵, Yijun Ruan^{5,6} and Dariusz Plewczynski^{1,2,5,*}

¹Centre of New Technologies, University of Warsaw, Warsaw 02-097, Poland, ²Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw 00-662, Poland, ³Centre for Bioinformatics and Data Analysis, Medical University of Białystok, Białystok 15-089, Poland, ⁴I-BioStat, Hasselt University, 3500 Hasselt, Belgium, ⁵The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA and ⁶Department of Genetics and Genome Sciences, UConn Health, Farmington, CT 06030-6403, USA

Received February 28, 2020; Revised May 02, 2020; Editorial Decision May 04, 2020; Accepted May 05, 2020

ABSTRACT

Structural variants (SVs) that alter DNA sequence emerge as a driving force involved in the reorganisation of DNA spatial folding, thus affecting gene transcription. In this work, we describe an improved version of our integrated web service for structural modeling of three-dimensional genome (3D-GNOME), which now incorporates all types of SVs to model changes to the reference 3D conformation of chromatin. In 3D-GNOME 2.0, the default reference 3D genome structure is generated using ChIA-PET data from the GM12878 cell line and SVs data are sourced from the population-scale catalogue of SVs identified by the 1000 Genomes Consortium. However, users may also submit their own structural data to set a customized reference genome structure, and/or a custom input list of SVs. 3D-GNOME 2.0 provides novel tools to inspect, visualize and compare 3D models for regions that differ in terms of their linear genomic sequence. Contact diagrams are displayed to compare the reference 3D structure with the one altered by SVs. In our opinion, 3D-GNOME 2.0 is a unique online tool for modeling and analyzing conformational changes to the human genome induced by SVs across populations. It can be freely accessed at <https://3dgenome.cent.uw.edu.pl/>.

INTRODUCTION

There is a plethora of evidence to be found in the literature for the significant role of genome spatial organization

in gene regulation for both health and disease (1–3). Structural variation (SV), encompassing deletions, duplications, insertions and inversions, is the main source of genetic variation in humans and it is shown to have a critical impact on higher-order chromatin conformation and gene functioning (1,4). Deletions, duplications, insertions and their combination may reorganize chromatin interactions by altering the DNA segments that are involved in the establishment of three-dimensional (3D) contacts. Inversions, on the other hand, may alter the directionality of the binding motifs of the CTCF proteins that bring the mentioned above segments together. SVs that do not overlap any interacting sites do not affect the resulting 3D structure (the number and relative arrangement of loops in the model), but they still contribute to the shortening or extending of chromatin loops.

By the effort of the 1000 Genomes Consortium, a catalogue of SVs identified in over 2500 human samples from 26 populations was created (5). Several examples of SVs present in non-coding regions were already reported to disrupt local 3D genome organization leading to an altered gene transcription (6,7) and in some cases causing disease (8–12). Nevertheless, this area of research is dominated by studies identifying spatial DNA structure for a selected and limited number of human cell lines (13,14). Thus, it lacks the broader perspective acquired by investigating the processes that shape genomic diversity in the human population as a whole - a focus of international studies such as the 1000 Genomes Project (1kGP) (5) or the Simons Genome Diversity Project (15). In this regard, the population analysis of SVs in the context of 3D genome structure can provide unique insights into biophysical mechanisms regulat-

*To whom correspondence should be addressed. Tel: +48 22 554 36 54; Fax: +48 22 554 08 01; Email: d.plewczynski@cent.uw.edu.pl

ing chromatin organisation and gene transcription at the population scale (4).

To address this issue, we have implemented our recently published SV-based modifying algorithm (4), adopted to generate altered chromatin interaction patterns, to our previously developed 3D genome modeling engine 1.0 (3D-GNOME) web server (16). This web service generates chromatin 3D structures using a Monte Carlo approach based on Chromatin Conformation Capture (3C) data, providing tools for their visualization and analysis. Users can generate 3D structures of a genomic region of interest by simply specifying its coordinates. Here we present the 2.0 version of 3D-GNOME. With this update, users gain the ability to predict SVs-driven changes in 3D conformations of specific loci in human genomes. In the default setting, 3D-GNOME 2.0 employs chromatin interaction paired-end tag sequencing (ChIA-PET) data for the GM12878 cell line (6) as the reference chromatin interaction map and sets of SVs emerging across human populations constructed by 1kGP as the source of genetic variation (18). When relying on our precomputed reference 3D model, users may also provide a particular individual's SVs. 3D-GNOME 2.0 returns both the reference 3D structure and the structure altered by SVs. Alternatively, users can submit their list of loops obtained from 3C based methods such as Hi-C or ChIA-PET and a custom SV list and generate 3D structures of their loci and genomes of choice (Figure 1).

3D-GNOME 2.0 provides novel tools to inspect, visualize and analyze 3D models of chromatin regions modified by SVs in the samples of interests at different levels of spatial chromatin organisation, starting with individual loops, genomic domains (chromatin contact domains-CCDs) and

continuing to full chromosomes. Differences between the reference and altered structure can be analyzed with diagrams of contacts, statistical plots, and three-dimensional models. According to our knowledge, this is the first easily accessible online tool for modeling conformational changes in the human genome induced by SVs in different populations. A schematic representation of the workflow of 3D-GNOME 2.0 is presented in Figure 1.

NEW FEATURES AND UPDATES

Basic web server architecture with extensions

New features introduced to 3D-GNOME consist of tools for both the processing and analysis of SVs associated data. The overall architecture of the web server is maintained. In detail, 3D-GNOME 2.0 has been implemented by using the WSGI application Flask framework (<https://palletsprojects.com/p/flask/>), and upgraded to be compatible with Python 3.6+. A validated request, potentially including SVs information, is saved to the MySQL database. The data processing pipeline is written in Python, together with external scripts written in C++, PHP and R. The web server generates contact diagrams, plots, statistics and 3D models. To view the models, we maintain an interactive viewer implemented in WebGL (<https://www.khronos.org/webgl/>), developed as part of the previous version of 3D-GNOME.

An important change in version 2.0 is the implementation of a method for recovering individualized genomic interaction patterns based on reference interaction patterns and an individual set of SVs published previously (4). It is also available as a separate python script at https://bitbucket.org/4dnucleome/spatial_chromatin_architecture/.

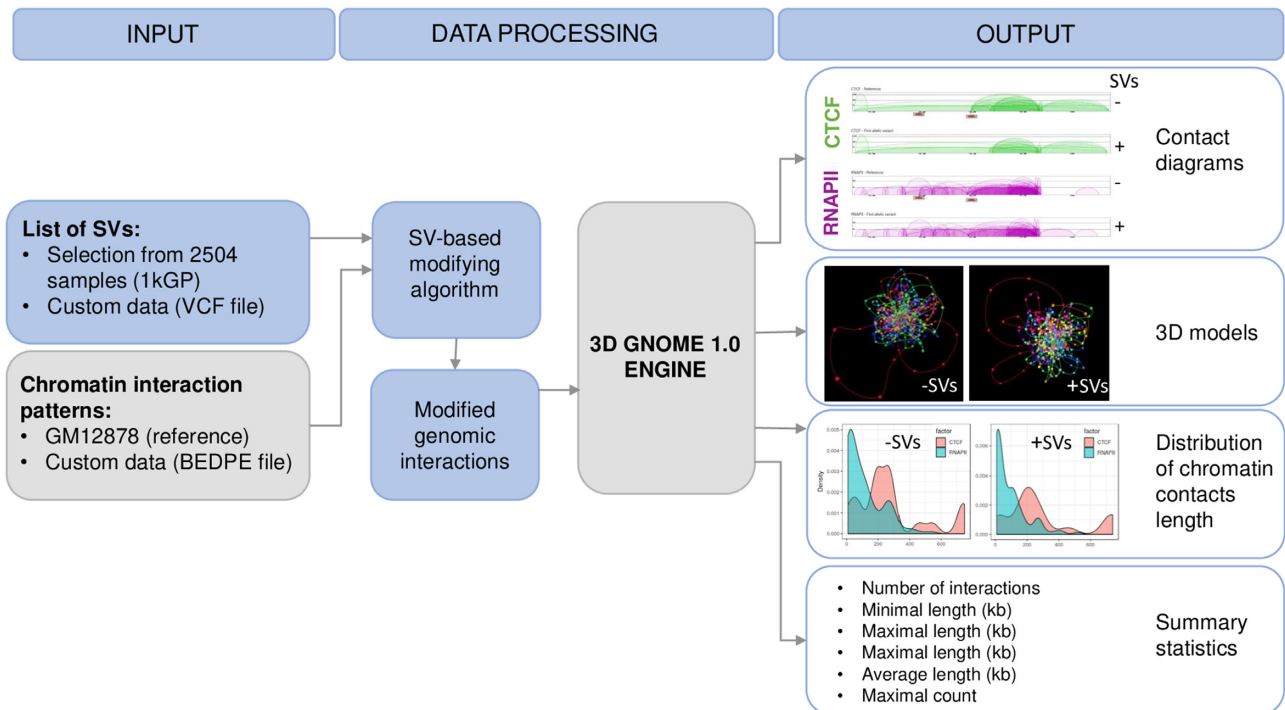


Figure 1. A schematic representation of the workflow of 3D-GNOME 2.0.

Modeling changes of spatial chromatin architecture induced by structural variants

We developed an algorithm for modeling changes of the genome topology by removing and creating contacts between chromatin interaction anchors based on genetic alterations introduced by SVs. The algorithm uses high-quality CTCF or RNA Polymerase II (RNAPII) ChIA-PET data collected for the GM12878 cell line as a reference chromatin interaction pattern (17). When given a set of SVs present in other lymphoblastoid genomes, it applies a series of simple rules to recover their individualized chromatin interaction patterns from the reference data. The resulting genomic interactions are then passed to the 3D-GNOME modeling engine to build 3D models of individualized chromatin structures. Of note, while GM12878 ChIA-PET data is set as the reference for modeling 3D genomes of the samples genotyped by the 1000 Genomes Consortium, in principle any genomic interaction data can be used as the reference. The method acts on the following types of SVs: deletions (DEL), duplications (DUP), insertions (INS) and inversions (INV). The algorithm's behavior as a function of the input SV type

is described below and represented schematically in Figure 2. The method was described in detail previously (4).

Deletions. Deletion removes all anchors within its scope, and therefore all the interactions they form with other anchors. Loops directly adjacent to the introduced deletion are elongated or shortened depending on the interaction pattern and size of the deletion. A deletion, which partially or completely overlaps the outermost anchor of a particular CCD, removes its boundary and merges it with the genomic region on the other side of the boundary.

Duplications and multiallelic copy number variants. Interactions that reside entirely in the duplicated region are repeated as a whole and introduced adjacently downstream in the genomic sequence. Similarly, a new copy of an anchor is created downstream from the original anchor. The original anchor maintains only the upstream interactions it formed before being duplicated. The downstream interactions are in turn connected to the duplicated anchor instead of the original one. Regarding CTCF mediated interactions,

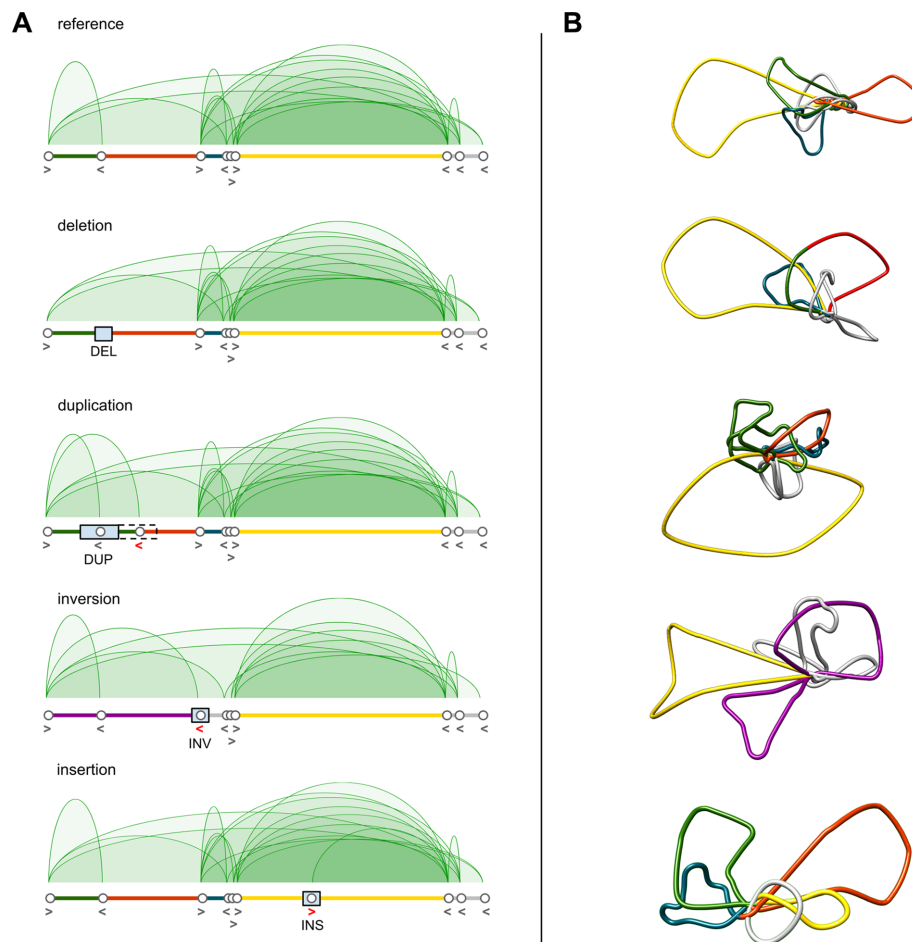


Figure 2. Schemes representing the behaviour of the computational algorithm implemented in 3D-GNOME 2.0 for prediction of changes in 3D chromatin contacts induced by SVs. (A) CTCF ChIA-PET contact diagrams for exemplary region *chr1:47656996-48192898* containing *TAL1* locus for the reference genome (GM12878) and upon introduction of SVs. Alteration of CTCF-mediated contact patterns upon addition of DUP, DEL, INV or INS to the genomic sequence is shown. SVs are marked as blue rectangles. CTCF anchors and their directionality are represented as white circles and arrows, correspondingly. Red arrows represent CTCF anchors' alterations induced by SVs. (B) 3D models of CTCF mediated chromatin structures corresponding to genomic regions shown in (A). Loops are coloured as genomic regions represented below CTCF contact diagrams depicted in (A).

if the original anchor is lacking downstream contacts, naturally there are no interactions that the duplicated anchor could take over. In such cases, the algorithm finds the closest CTCF anchor with which it can create a convergent loop and connects them. This strategy reflects the currently assumed mechanism of CTCF-mediated chromatin loop formation, known as extrusion (19). According to this model, chromatin is pulled through the ring-shaped cohesin complex until the cohesin ring stops at an obstacle larger than the ring lumen. In this model, these obstacles correspond to CTCF proteins bound to the DNA motifs on the opposite DNA strands. Furthermore, if a duplication expands over a CCD boundary, parts of the CCDs, placed at the breakpoint after duplication, are fused. On top of having a direct impact on interacting loci, duplications can also increase the lengths of chromatin-loop spanning the duplicated regions. The effects of Multiallelic Copy Number Variants (mCNVs) are introduced in the 3D genome by performing multiple duplications or deletions.

Insertions. Inserted sequences are scanned for CTCF motifs using the PFM matrix. If a motif is found, the algorithm treats it as a new CTCF anchor and finds the closest anchor with which it can create a convergent loop.

Inversion. The algorithm reverses the directionality of the anchors affected by the inversion and removes all original contacts they formed. Next, it matches each of these anchors with the closest anchor of opposite orientation to create a convergent loop. If the considered anchors correspond to protein factors that do not bind any specific DNA motifs or if the orientation of the motifs they bind is irrelevant for the loop formation, the algorithm links such anchors with the closest ones, regardless of their directionality. If as a result of inversion, the algorithm links anchors from adjacent CCDs, the boundary between them is removed and the CCDs are merged.

To summarize, if as a result of the introduction of a given SV, an anchor loses all its interactions, it is joined with the nearest anchor of the opposite orientation (in the case of CTCF-mediated interactions), or to the nearest anchor without additional criteria. If an SV does not intersect any CTCF binding sites (or correspondingly, any other protein binding site, like RNAPII), no changes will be introduced to the chromatin interaction pattern, except for the change in length of certain loops.

The algorithm introduces changes in PET clusters, leaving singletons unchanged, thus, singleton data in sample variants will be equivalent to their reference singleton sets.

Input

The 3D-GNOME 2.0 web server is able to model 3D structures across individuals. In particular, one can model conformation of the genomes genotyped by the 1000 Genomes Consortium, and study topological genome variability in the human population. Users can choose to model chromatin structures using CTCF interactions only or CTCF and RNAPII interactions at the same time. The only input information that needs to be provided is the location of

the genomic region of interest and the list of SVs identified in a given genome. The former could be solely specified by chromosome number and coordinates, for example, *chr14:35138000-36160000* (GM12878 ChIA-PET data will be used as the reference in this case), or the user can upload a BEDPE-format file with their own 3D chromatin interaction map data containing locations and frequencies of long-range contacts (obtained from 3C based methods such as Hi-C or ChIA-PET). The latter can be provided either as a custom list of SVs in the VCF format or by choosing SVs of interest from the predefined list of IDs identical to the ones from the 1kGP. The VCF file can be uploaded using the *Upload VCF file* option in the field '*List of structural variants*'. If more than one sample is to be analysed, sample IDs should also be entered, separated by commas into the text box '*IDs of selected samples*'. Alternatively, users can select IDs of 1kGP samples by choosing the '*Select Sample IDs*' option in the '*List of structural variants*' field.

To reduce calculation time, we cached interaction data for 2,504 genomes in both allelic variants. The '*Submit*' button starts a simulation and a URL, pointing to the results page, appears. During the computations, the results page is constantly refreshed until the task is completed, after which the results are displayed.

Output

In 3D-GNOME 2.0 we extended the user interface by adding a menu on the left side contains a list of checkboxes that allow users to show or hide selected results such as contact diagrams, singleton heatmaps, data statistics, and plots. The desired result details may be displayed for chosen genomic samples, for both allelic variants separately. The selected elements are shown in the centre of the screen.

The main improvement in the output content includes contact diagrams that allow users to compare changes in interaction patterns introduced by SVs of interest. SVs emerging in a selected region are annotated on a reference contact diagram by labels and lines colored according to the SV type. Two separate diagrams of chromatin contacts are displayed for each reference and variant case, one representing genomic interactions mediated by CTCF, another showing RNAPII-mediated contacts (Figure 3A). Next, all statistics, such as Number of Interactions, Minimal/Maximal/Average Length of interactions and Maximal Count Frequency of the CTCF, RNAPII and singleton interactions are calculated and presented in separate plots for each variant. Additionally, a plot showing the distribution of interaction lengths is generated for each sample (Figure 3B). This enables a convenient comparison between the reference and the variant. Interaction sets displayed on the charts may be downloaded in the BEDPE format.

The menu contains URLs that link to the page containing an interactive 3D viewer with the pre-loaded chromatin model of the selected structure. 3D-GNOME 2.0 extends the possibilities of the 3D viewer with the ability to display both 3D models of the reference structure and the variant side by side (Figure 3C, D). Additionally, the 3D-Viewer Control Menu has the option to save the 3D model locally in PDB (Protein Data Bank) or XYZ format. This allows lo-

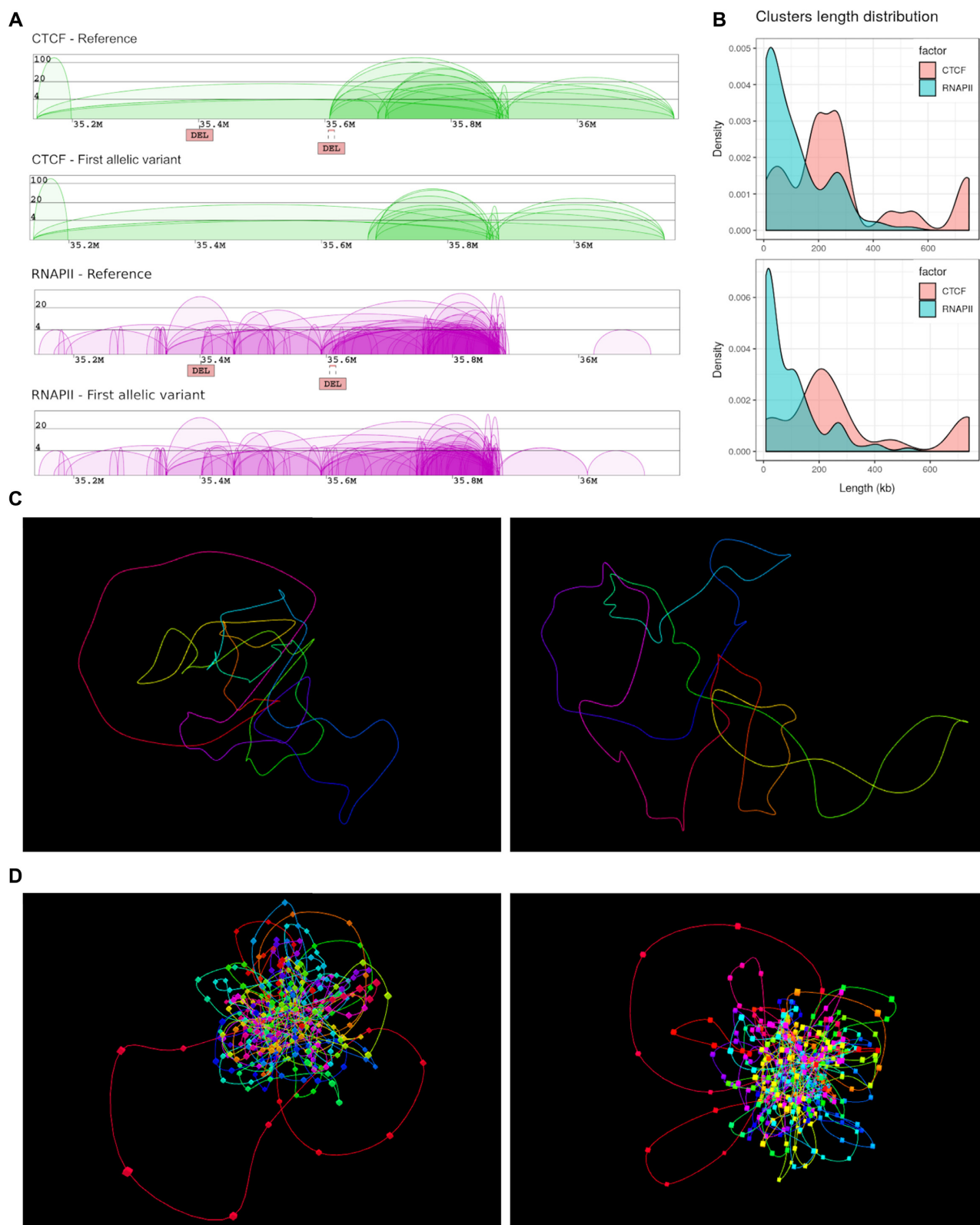


Figure 3. Output of 3D-GNOME 2.0, for an exemplary region (*chr14:35138000-36160000*) affected by two deletions: (*chr14:35401403-35401724* and *chr14:35605439-35615196*). (A) Screenshot of the result page with diagrams of chromatin contacts mediated by CTCF (green) and RNAPII (purple) for both the reference genome (based on GM12878 data) and the variant genome (based on SVs from HG00099); the deletion *chr14:35605439-35615196* (right DEL) disrupts CTCF and RNAPII interactions. (B) Clusters length distribution for CTCF and RNAPII protein factors for reference genome (top) and HG00099 (bottom). (C, D) Representation of 3D models in the 3D viewer, proposed using only CTCF interactions (panel C) or both CTCF and RNAPII data (panel D) from the reference genome (left) and that affected by SVs (right).

cal examination of the model with a molecular visualization software such as UCSF Chimera (<https://www.cgl.ucsf.edu/chimera/>). It should be highlighted that taking into consideration the nature of 3C experiments, which are performed using millions of cells, these models represent the average of structures existing in the population of the cells rather than structures from individual cells. Several successful analyses have been, however, already published using the former kind of data. For example, as described in (4), analyzing the distribution of distances between specific genomic elements, like enhancers and promoters, and gene expression profiles in a given region can lead to uncovering meaningful associations.

Computational time of modeling depends on different parameters, such as: the size of the region, the number of interactions-mediated proteins (CTCF only or CTCF and RNAPII together) or whether the user selects/provides the list of SVs or not. For example, for a region *chr14:35138000-36160000* computations time is (I) ~30 s if only CTCF interactions are considered for modeling, (II) ~2 min if both CTCF and RNAPII interactions are used, and (III) ~8 min. if both CTCF and RNAPII interactions are used and the sample is modified by SVs from the HG00099 cell line.

CONCLUSIONS AND FUTURE PLANS

3D-GNOME 2.0 provides users with a new capacity to process structural variation data. Novel features enable users to model changes in chromatin contacts caused by SVs and allow to compare these changes using diagrams of contacts, statistical plots and 3D models. We believe that 3D-GNOME 2.0 is an easily accessible tool, valuable to scientists who wish to study processes that shape 3D chromatin architecture in the nucleus, specifically from a population perspective. In the future, we plan to implement a GPU-accelerated version of our modeling algorithm. This will help reduce the computational time and therefore provide a more effective way to analyse large population datasets. Furthermore, we also aim to improve 3D visualization and model analysis.

ACKNOWLEDGEMENTS

We thank Veronika Mancikova for language editing and proofreading of the manuscript and Agnieszka Bucka and Agnes Alcantara Paculdar for critical reading of the text.

FUNDING

Polish National Science Centre [2014/15/B/ST6/05082]; Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (TEAM to D.P.); 'Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation' within the 4DNucleome National Institute of Health program, and by the European Commission as European Cooperation in Science and Technology COST actions: CA18127 'International Nucleome Consortium' (INC) [IU54DK107967-01]; 'Impact of Nuclear Domains On Gene Expression and Plant Traits' [CA16212]; Horizon 2020 Marie Skłodowska-Curie ITN Enhpathy grant

'Molecular Basis of Human enhanceropathies'; D.P. and M.W. were supported by the RENOIR Project from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie [691152]; Ministry of Science and Higher Education (Poland) [W34/H2020/2016, 329025/PnH/2016]. Funding for open access charge: Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (TEAM to DP).

Conflict of interest statement. None declared.

REFERENCES

1. Spielmann, M., Lupianez, D.G. and Mundlos, S. (2018) Structural variation in the 3D genome. *Nat. Rev. Genet.*, **19**, 453–467.
2. Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A. *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (New York, N.Y.)*, **351**, 1454–1458.
3. Schoenfelder, S. and Fraser, P. (2019) Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.*, **20**, 437–455.
4. Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y. and Plewczynski, D. (2019) Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome Biol.*, **20**, 148.
5. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
6. Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Włodarczyk, J., Ruszczycy, B. *et al.* (2015) CTCF-Mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
7. Heinz, S., Texari, L., Hayes, M.G.B., Urbanowski, M., Chang, M.W., Givarkes, N., Rialdi, A., White, K.M., Albrecht, R.A., Pache, L. *et al.* (2018) Transcription elongation can affect genome 3D structure. *Cell*, **174**, 1522–1536.
8. Lupianez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.
9. Weischenfeldt, J., Dubash, T., Drains, A.P., Mardin, B.R., Chen, Y., Stutz, A.M., Waszak, S.M., Bosco, G., Halvorsen, A.R., Raeder, B. *et al.* (2017) Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.*, **49**, 65–74.
10. Kantidze, O.L., Gurova, K.V., Studitsky, V.M. and Razin, S.V. (2020) The 3D genome as a target for anticancer therapy. *Trends Mol. Med.*, **26**, 141–149.
11. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.)*, **337**, 1190–1195.
12. Talkowski, M.E., Mullegama, S.V., Rosenfeld, J.A., van Bon, B.W., Shen, Y., Repnikova, E.A., Gastier-Foster, J., Thrush, D.L., Kathiresan, S., Ruderfer, D.M. *et al.* (2011) Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. *Am. J. Hum. Genet.*, **89**, 551–563.
13. 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**, 1665–1680.
14. Szalaj, P., Tang, Z., Michalski, P., Pietal, M.J., Luo, O.J., Sadowski, M., Li, X., Radew, K., Ruan, Y. and Plewczynski, D. (2016) An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization. *Genome Res.*, **26**, 1697–1709.
15. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A. *et al.* (2016) The

- simons genome diversity Project: 300 genomes from 142 diverse populations. *Nature*, **538**, 201–206.
16. Szalaj,P., Michalski,P.J., Wroblewski,P., Tang,Z., Kadlof,M., Mazzocco,G., Ruan,Y. and Plewczynski,D. (2016) 3D-GNOME: an integrated web service for structural modeling of the 3D genome. *Nucleic Acids Res.*, **44**, W288–W293.
17. Fullwood,M.J. and Ruan,Y. (2009) ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.*, **107**, 30–39.
18. Chaisson,M.J.P., Sanders,A.D., Zhao,X., Malhotra,A., Porubsky,D., Rausch,T., Gardner,E.J., Rodriguez,O.L., Guo,L., Collins,R.L. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, **10**, 1784.
19. 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 (New York, N. Y.)*, **326**, 289–293.

Structural Bioinformatics

cudaMMC - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling

Michał Wlasnowolski^{1,2,†}, Paweł Grabowski^{1,†}, Damian Roszczyk^{1,†}, Krzysztof Kaczmarski¹ and Dariusz Plewczynski^{1, 2,*}

¹Faculty of Mathematics and Information Science, Warsaw University of Technology, ²Centre of New Technologies, University of Warsaw

† equally contributed

*To whom correspondence should be addressed.

Abstract

Motivation: Investigating the 3D structure of chromatin provides new insights into transcriptional regulation. With the advancements in 3C new-sequencing techniques such as ChiA-PET and Hi-C, there has been a substantial increase in the volume of collected data, necessitating faster algorithms for chromatin spatial modelling. This study presents the cudaMMC method, which utilises the Simulated Annealing Monte Carlo approach extended by GPU-accelerated computing to generate ensembles of chromatin 3D structures efficiently.

Results: The *cudaMMC* calculations demonstrate significantly faster performance and lower (better) model scores compared to our previous method on the same workstation. *cudaMMC* substantially reduces the computation time required for generating ensembles of large chromatin models, making it an invaluable tool for studying chromatin spatial conformation.

Availability: Open-source software and manual and sample data are freely available on <https://github.com/SFGLab/cudaMMC>

Contact: Dariusz.Plewczynski@pw.edu.pl

1 Introduction

In recent years, the development of high-throughput sequencing methods and chromosome conformation capture (3C) technology has shown the significant influence of chromatin spatial conformation on genetic expression. Dynamic changes in the 3D structure at various levels of DNA spatial organisation: from the entire chromosomes, chromosomal territories, domains (TAD, CCDs), or single chromatin loops, affect the transcription level of individual genes (reviewed in (Chiliński *et al.*, 2021)). There is plenty of evidence that rearrangements of the spatial chromatin structure could alter gene expression. These changes may occur dynamically in the cell environment induced by heat stress or cell differentiation (Ray *et al.*, 2019, Pei *et al.*, 2020), as well as caused by genetic mutations, viruses infections, Structural Variations and DNA methylations (Sadowski *et al.*, 2019, Lazniewski *et al.*, 2019, Heinz *et al.*, 2019). To investigate this phenomena, various algorithms for generating chromatin 3D structures were developed (Kadluf *et al.*, 2020, Di Pierro *et al.*, 2016), among others one is the *3D-GNOME* approach (Szałaj *et al.*, 2016). *3D-GNOME* allows for generating ensembles of the 3D models of DNA using the simulated annealing Monte Carlo approach based on a map of chromatin contacts mediated by specific proteins like CTCF or RNAPII that play a crucial role in chromatin spatial organisation. This modelling

technique considers the DNA hierarchical structure organisation, starting from chromosome positioning through chromatin domains (TADs, CCDs) to a single chromatin loop shape. Due to the substantial development of 3C techniques, the data volumes of the chromatin contacts have increased significantly. This substantially increases the computation time needed to model the ensemble of large chromosomal structures, making it less useful for genome-wide modelling. To address this issue, we developed *cudaMMC*, which extended 3D-GNOME by implementing parallel acceleration using GPUs, resulting in a significant speed-up of calculations while maintaining modelling quality. This is necessary for calculation of spatial distance distribution between specific genomic loci, for which purpose we applied *cudaMMC* recently in the 3D-GNOME web server update (Wlasnowolski *et al.* 2020) for statistical analysis of changes of distances between gene promoters and enhancers. We also added a new option for the output file format, mmCIF, which enables the presentation and analysis of chromatin 3D models using commonly used molecular 3D viewers, such as UCSF Chimera (Pettersen *et al.*, 2004).

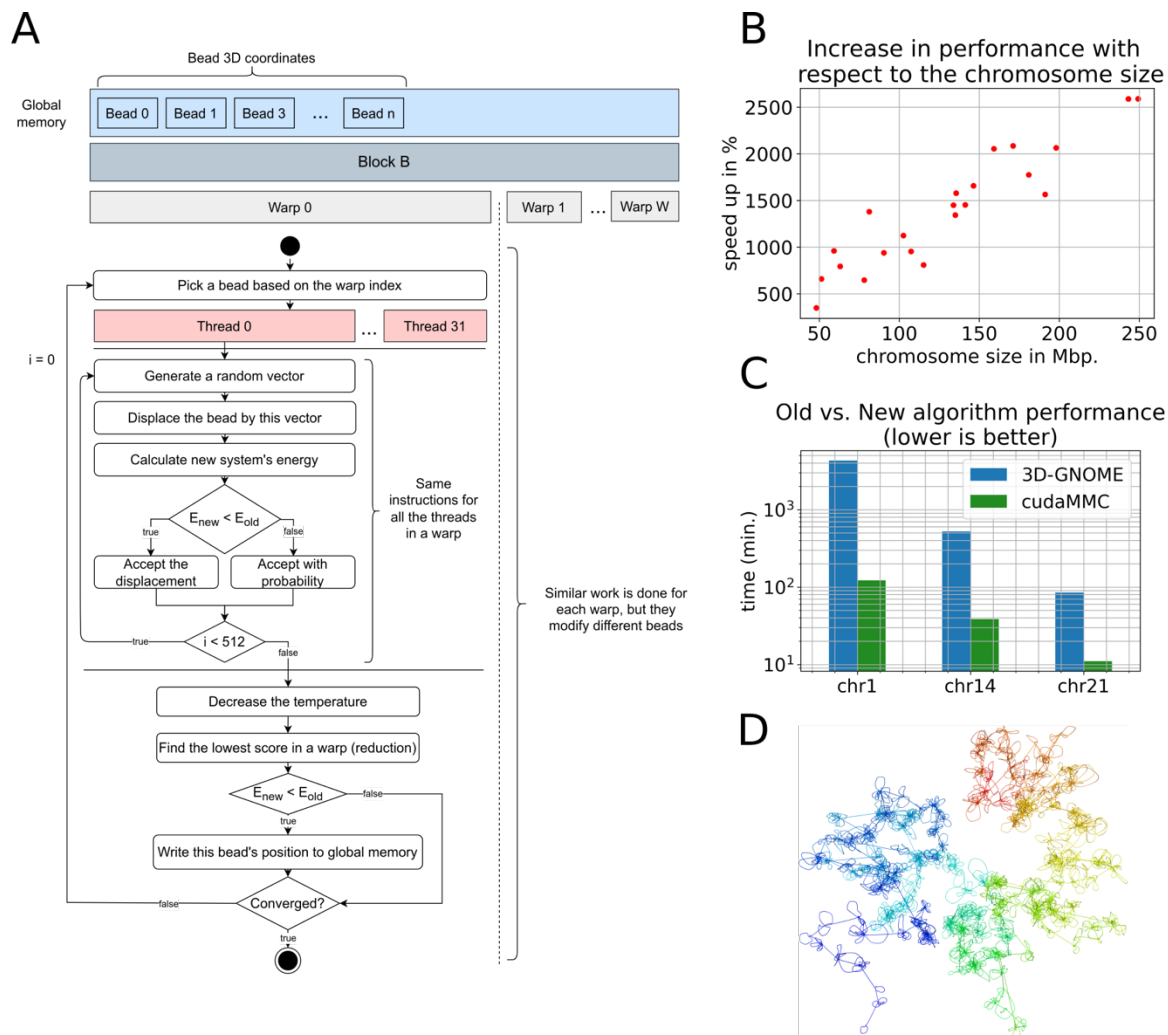


Fig.1. (A) Pipeline of the Parallel Simulated Annealing CUDA-Based Algorithm for Chromatin 3D Structure Modeling. (B) Speedup of the parallel algorithm with respect to the chromosome size. (C) Comparison of the performance between the 3D-GNOME (old) and cudaMMC (new) algorithms based on generating ensembles of 100 models for selected chromosomes. (D) Full chromosomal chromatin 3D structure of the chr1 based on CTCF ChIA-PET chromatin interactions for the GM12878 cell line mapped on GRCh37.

2 Materials and methods

2.1 3D-GNOME - CPU-oriented approach

The 3D-GNOME method consists of two stages. In the first one, a chromosome is divided into several regions based on PET cluster interaction patterns so that each region can correspond to one topological domain. Next, Monte Carlo simulated annealing is used to position beads to minimise energy function, considering the distance between beads corresponding to different regions. The second stage models the position of chromatin anchors within each domain independently based on energy terms. Next, chromatin loops are modelled by inserting sub-anchor beads between adjacent anchors, wherefore their position and shape are set using minimised energy function again.

2.2 cudaMMC approach GPU-extended

A successful GPU algorithm must find a balance between local and global operations to minimise synchronisation and maximise parallel computations. GPU processing is based on so-called warps: synchronised cooperative groups of 32 threads, which may perform partially independent steps and exchange data using intrinsic commands or highly specialised shared memory. Our implementation of parallel simulated annealing based on the Monte Carlo method uses warps to perform random moving of the beads (Fig. 1A). A single warp optimises a single bead using 32 threads to try random moves of that bead iteratively. If some position is better, it is stored for further evaluation. Finally, a warp selects a minimum from its threads using parallel reduction. This strategy allows for the parallel fitting of thousands of beads in 32 different directions each.

2.3 Chromatin interaction data

We compared cudaMMC and 3D-GNOME algorithms on long-read ChIA-PET CTCF chromatin interaction data for the GM12878 cell line mapped

cudaMMC approach

on GRCh37, for which data 3D-GNOME was designed. We have also performed modelling tests using cudaMMC on *in situ* ChIA-PET data for the GM12878 cell line mapped on GRCh38. We omitted the 3D-GNOME performance testing on GRCh38 data because of the excessive computational time required caused by the massive increase of data gained by the new ChIA-PET method.

3 Comparison of performance between cudaMMC and 3D-GNOME

We compared the performance of the cudaMMC and 3D-GNOME methods on a workstation with a NVIDIA Pascal GPU architecture and an Intel Core i9-7920X CPU. The comparison was based on data sizes of varying sizes. The cudaMMC algorithm resulted in a speed-up of 3x to 25x for a single chromosome modelling, depending on the chromosome size (Figure 1B). The biggest advantage of the algorithm speed-up was observed for generating ensembles of models. Using GRCh37 data, we performed ensembles consisting of 100 models each using both methods. The cudaMMC algorithm decreased computation time from ~85 to ~11 min. for chr21, from ~8.5h to ~38 min. for chr14, and from ~3 days to 2h for chr1. Moreover, we tested the performance of the cudaMMC algorithm for whole chromosomal modelling based on *in situ* ChIA-PET data mapped on GRCh38. The ensemble of 100 models generated for chr1 took ~8h, for chr14 ~2h 20 min., and for chr21 ~21 min. (Figure 1C). These results were obtained on Pascal architecture, but cudaMMC might also be configured on Turing or Ampere NVIDIA devices perhaps with even better results. We added a tool for converting output into mmCIF format to make it more accessible for common usage. We chose this format instead of PDB because it can represent a higher number of beads for one structure, which is necessary for high-resolution whole chromosome models (Figure 1D). Overall, our results show that the cudaMMC algorithm can significantly improve performance for 3D modelling of chromatin structures, making it a valuable tool for researchers in this field.

4 Conclusions

In the cudaMMC method, parallelising calculations on the GPU enabled a massive reduction in computation time compared to the previous CPU-oriented approach of 3D-GNOME. This speed-up in calculations enables the generation of a high number of model ensembles in a reasonable time frame, thereby offering an opportunity for statistical analysis. For example, as part of the update to the 3D-GNOME web server, we have recently added a tool for statistical analysis of ensembles, which calculates changes in distances between gene promoters and enhancers of two model ensembles that differ in chromatin loop patterns caused by Structure Variants. To facilitate this, we have set up the cudaMMC software on the NVIDIA DGX A100 cluster, enabling the running of multiple modelling tasks simultaneously on several GPU cards and facilitating analysis based on new, large datasets of spatial chromatin data. We believe that cudaMMC, the GPU-accelerated 3D-GNOME modelling engine, will significantly enhance scientists' investigations into various aspects of chromatin 3D structures.

Funding

This work has been supported by National Science Centre, Poland (2019/35/O/ST6/02484 and 2020/37/B/NZ2/03757), and European Commission Horizon 2020 Marie Skłodowska-Curie ITN Enhpathy grant 'Molecular Basis of Human enhanceropathies'. Research was co-funded by Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme. Computations were performed thanks to the Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology using Artificial Intelligence HPC platform financed by Polish Ministry of Science and Higher Education (decision no. 7054/IA/SP/2020 of 2020-08-28).

Conflict of Interest: none declared.

References

- Chiliński, M., Sengupta, K., & Plewczynski, D. (2021, August). From DNA human sequence to the chromatin higher order organisation and its biological meaning: Using biomolecular interaction networks to understand the influence of structural variation on spatial genome organisation and its functional effect. In *Seminars in Cell & Developmental Biology*. Academic Press.
- Ray, J., Munn, P. R., Vihervaara, A., Lewis, J. J., Ozer, A., Danko, C. G., & Lis, J. T. (2019). Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proceedings of the National Academy of Sciences*, 116(39), 19431-19439.
- Pei, L., Huang, X., Liu, Z., Tian, X., You, J., Li, J., ... & Wang, M. (2022). Dynamic 3D genome architecture of cotton fiber reveals subgenome-coordinated chromatin topology for 4-staged single-cell differentiation. *Genome biology*, 23(1), 1-25.
- Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y., & Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome biology*, 20(1), 1-27.
- Lazniewski, M., Dawson, W. K., Rusek, A. M., & Plewczynski, D. (2019, June). One protein to rule them all: the role of CCCTC-binding factor in shaping human genome in health and disease. In *Seminars in cell & developmental biology* (Vol. 90, pp. 114-127). Academic Press.
- Heinz, S., Texari, L., Hayes, M. G., Urbanowski, M., Chang, M. W., Givarkes, N., ... & Benner, C. (2018). Transcription elongation can affect genome 3D structure. *Cell*, 174(6), 1522-1536.
- Kadluf, M., Rozycka, J., & Plewczynski, D. (2020). Spring Model—chromatin modeling tool based on OpenMM. *Methods*, 181, 62-69.
- Szalaj, P., Tang, Z., Michalski, P., Pietal, M. J., Luo, O. J., Sadowski, M., ... & Plewczynski, D. (2016). An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization. *Genome research*, 26(12), 1697-1709.
- Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G., & Onuchic, J. N. (2016). Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences*, 113(43), 12168-12173.
- Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., ... & Plewczynski, D. (2020). 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Research*, 48(W1), W170-W176.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13), 1605-1612.

3D-GNOME 3.0: a three-dimensional genome modelling engine for analysing changes of promoter-enhancer contacts in the human genome

Michał Wlasnowolski^{1,2}, Michał Kadlof¹, Kaustav Sengupta^{1,2} and Dariusz Plewczynski^{1,2,*}

¹Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, 00-662, Poland and

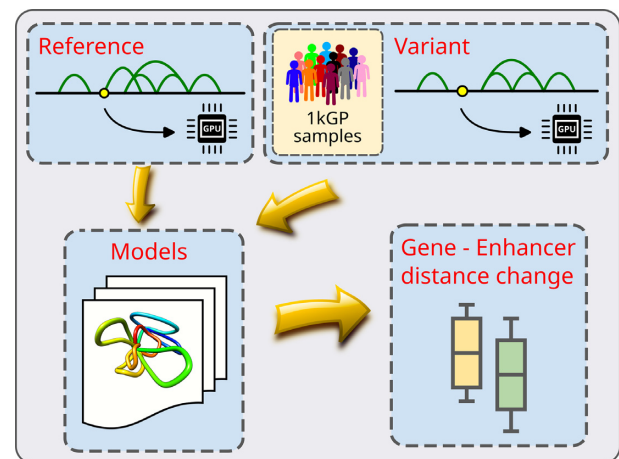
²Centre of New Technologies, University of Warsaw, Warsaw, 02-097, Poland

Received February 23, 2023; Revised April 14, 2023; Editorial Decision April 22, 2023

ABSTRACT

In the current update, we added a feature for analysing changes in spatial distances between promoters and enhancers in chromatin 3D model ensembles. We updated our datasets by the novel *in situ* CTCF and RNAPII ChIA-PET chromatin loops obtained from the GM12878 cell line mapped to the GRCh38 genome assembly and extended the 1000 Genomes SVs dataset. To handle the new datasets, we applied GPU acceleration for the modelling engine, which gives a speed-up of 30× versus the previous versions. To improve visualisation and data analysis, we embedded the IGV tool for viewing ChIA-PET arcs with additional genes and SVs annotations. For 3D model visualisation, we added a new viewer: NGL, where we provided colouring by gene and enhancer location. The models are downloadable in mmCIF and xyz format. The web server is hosted and performs calculations on DGX A100 GPU servers that provide optimal performance with multitasking. 3D-GNOME 3.0 web server provides unique insights into the topological mechanism of human variations at the population scale with high speed-up and is freely available at <https://3dgnome.mini.pw.edu.pl/>.

GRAPHICAL ABSTRACT



INTRODUCTION

One of the primary challenges in human genetics, precision medicine, and evolutionary biology is deciphering gene expression regulation and understanding the transcriptional effects of genome variation (1–3). The three-dimensional organisation of chromatin and the spatial proximity between enhancers and gene promoters have been shown to impact gene expression significantly (4–7). Additionally, structural variants (SVs) that alter chromatin structure can profoundly affect gene regulation (8). Genomic studies indicate that SVs can directly impact the interactions between the promoter and enhancer regions of the chromatin (9,10), which could lead to the development of new therapeutic targets and diagnostic tools (11,12). Thus, developing an accessible tool to investigate these complex interactions is essential for advancing the understanding of gene expression regulation and genome variation. This paper proposes a new version of the 3D-GNOME web server (13,14) that provides tools for comparing different 3D structures of the genome (Figure 1). It enables the analysis of changes in the modelled distance distribution between enhancers and gene pro-

*To whom correspondence should be addressed. Tel: +48 222347219; Email: Dariusz.Plewczynski@pw.edu.pl

IGV (17) and the model viewer to NGL (18). Also, for better visualisation of the modelling results, models are coloured based on gene promoter body and enhancer location.

As far as we are aware, only a limited number of web servers are available that offer the ability to generate chromatin 3D models and detailed genomic feature analysis (19,20). However, none of these web servers provides the option to calculate changes in spatial distances between enhancers and genes caused by structural variants in different human populations by generating full ensembles of chromatin 3D models based on high-resolution ChIA-PET data, which is why we find our new feature unique. This new release gives abilities for analysing the potential impact of genome spatial changes on gene activity, allowing for a deeper understanding of gene regulation and cellular processes.

NEW FEATURES AND UPDATES

New datasets

In the previous version of 3D-GNOME web server, the modelling of chromatin structure was based on long-range ChIA-PET data, including CTCF and RNAPII chromatin interactions of the GM12878 cell line mapped onto the GRCh37 reference genome, as well as structural variants from 2502 samples from the 1000 Genomes Project release 3, also mapped onto GRCh37.

In the current version, we have replaced the previous dataset of CTCF and RNAPII interactions in the GM12878 cell line, which was obtained from *long-read* ChIA-PET (21,22), with high-resolution data from *in situ* ChIA-PET (23), which was mapped onto the GRCh38 reference genome. The new dataset provides substantially more chromatin interactions with higher confidence and offers a more comprehensive and accurate view of the genome's spatial architecture in the GM12878 cell line. As a result of this new dataset, the quality of chromatin 3D models generated using 3D-GNOME has also improved.

The structural variant dataset (15) has been updated, with the previous GRCh37 version replaced with a GRCh38 version. The number of samples expanded to 3202 by including 30x high-coverage data from the NYGC on GRCh38. These updates provide a more comprehensive and accurate representation of chromatin structure, enabling further analysis and understanding of its impact on gene expression.

GPU-accelerated modelling engine

We have implemented GPU acceleration into our modelling engine, which is based on the Simulated Annealing Monte Carlo method, to address the significant increase in calculation time when analysing ensembles of chromatin 3D models using much larger datasets of chromatin 3D contacts. As a result, we have achieved a 30x speed-up compared to the previous version. To facilitate subsequent analysis, we have converted the models from the hcm, 3D-GNOME native format to the XYZ and mmCIF formats, which can handle models with many more beads than the PDB format.

Updated web server architecture

The primary modelling task is performed on the Eden cluster, an in-house heterogeneous computing cluster equipped with Nvidia DGX A100 nodes. The Eden cluster is controlled by the Slurm (24) queuing system, which is deployed at the Faculty of Mathematics and Information Science at Warsaw University of Technology. The 3D-GNOME web interface runs on an LXC container in a ProxMox environment.

When a user submits a modelling request, the Flask web server executes a sequence of tasks, including validating the data, saving the data in a shared location with the Eden cluster, creating a database entry for the new task, and passing the task identifier to a concurrently running Gnu Parallel process (25). Gnu Parallel runs a Python script with a pipeline that performs local data pre-processing and then sends a request to run the modelling on the cluster.

Communication between the container and the Slurm controller is done through a REST API. The pipeline process periodically checks the status of the Slurm task, and when it receives information about the completion of the computation, it performs post-processing and updates the database entry. Once the modelling is complete, the user can view the results by refreshing the page.

Ensemble analysis

A key feature of the current update is the ability to analyse changes in spatial distances between gene promoters and enhancers caused by structural variants. This involves generating multiple chromatin 3D models for a specific chromatin region, both for the reference chromatin contact pattern and for the pattern affected by the SVs. Genes (GRCh38) and enhancers (based on Enhancer Atlas 2.0 (26), liftovered to GRCh38) are mapped onto each model, and the Euclidean distance between enhancers is calculated. The distance measure is specific to the 3D-GNOME engine, so the key factor for analysis is a change in distance distribution, as demonstrated in Sadowski *et al.* (27). To test the significance of the change in distance distribution, we use the Mann–Whitney *U* test with a *P*-value threshold of 0.05. This analysis provides insights into the impact of SVs on gene regulation by identifying changes in the spatial proximity between gene promoters and enhancers.

Input

In the request form, as in the previous version, the user may use prepared datasets for GM12878 chromatin interactions, set the region of interests and 3D modelling parameters and choose the sample ID of structure variation from the 1000 Genome Project database (15). It is also possible to upload chromatin interactions in BEDPE format or SVs in VCF format (VCFv4.2).

In the current version, we add to the form checkbox that runs ensemble analysis and sets the number of models in the ensemble.

Output

The 3D-GNOME web server presents new results in a fully responsive table and a boxplot generator for visualising the

distribution of gene-enhancer distances. In addition, the web server has been updated with new tools for data visualisation, building on the functionality of the previous version (Figure 2).

Promoter–enhancer distance comparison. We present the results of comparisons of promoter–enhancer distances in a responsive table generated using the Bootstrap package. The table displays the genes, gene types (including pseudogenes), enhancers with an enhancer score, average gene-enhancer distance in an ensemble in the reference and variant structures, as well as differences between these two ensembles and p-values of the significance of those differences. Users can search, sort, and filter the results by columns. Furthermore, we have added an option to generate a distribution boxplot for a selected region. The user may select rows with gene-enhancer pairs using checkboxes and use the ‘Generate distance boxplots’ button to submit the task. After that, using Ajax, the task is asynchronously transferred to Flask, the boxplots are calculated, and they are drawn using the Seaborn package (28).

Finally, after the automated page refreshing, the boxplots are viewed on the result web page. The boxplots with distances are displayed on the screen below the table and can also be downloaded from the download section.

Genome browser. We have integrated the Interactive Genome Viewer (IGV) (17) as a genome browser, providing an alternative to presenting arc diagrams in static PNG format as in the previous version. IGV is a highly responsive tool that allows users to visualise and manipulate genomic tracks, such as chromatin contact arcs, and gene, enhancer, and structural variation annotations for reference and variant samples. All data are displayed on the GRCh38 genome assembly. One notable feature of IGV is its ability to save results in vector file format (SVG), which makes it easy to present results outside the web server.

3D viewer. We have integrated NGL (18), a modern and interactive molecular visualisation tool, to present chromatin 3D structures dynamically and intuitively. NGL enables users to explore and interact with 3D models generated by 3D-GNOME, allowing them to adjust the view, zoom in and out, and rotate the structures to better understand the spatial relationships between genes and enhancers. Users can investigate the impact of structural variation on the 3D organisation of the genome by displaying both reference and variant 3D models on two separate 3D viewers, with coloured by genes and enhancers mapped on them.

Download section. The download section now includes the entire generated ensemble of models in mmCIF and XYZ format for manual analysis using common tools for visualising 3D structures, such as UCSC Chimera. Additionally, a *tsv* file with the results of the distance analysis is provided, including gene IDs, gene and enhancer coordinates, average distances in the ensemble, and the results of the Mann–Whitney *U* test of distribution changes (*P*-value and statistical value). Each gene-enhancer distance boxplot generated by clicking the ‘Generate boxplots’ button is also included in the output file folder.

CONCLUSIONS AND FUTURE PLANS

This latest update to 3D-GNOME web server provides an advanced tool for analysing modelled distance changes between enhancers and gene promoters. This is a valuable resource for exploring the impact of 3D chromatin structure on gene transcription and regulation. The new version offers significantly improved speed and efficiency due to GPU acceleration and Eden cluster architecture, enabling faster and more efficient chromatin modelling and analysis. We have also added new tools, including the NGL Viewer and IGV genome browser, which enhance the user experience by providing an intuitive and visually appealing way to analyse data.

In the near future, we plan to extend our datasets of chromatin interactions by including additional cell lines, such as H1ESC, HFFC6 and WTC11, as well as new structure variants from the Simons Diversity Projects for modern humans and archaic populations, such as Neanderthals and Denisovans. Including these archaic populations will provide a unique opportunity to investigate the evolution of chromatin structure and its impact on gene regulation across different populations, shedding new light on the history and diversity of our species. We also plan to add new input formats and datasets, such as Hi-C data, which are already standard in the scientific community. To facilitate this, we plan to implement in the web server chromatin loop calling software, which is necessary for converting native Hi-C data for 3D-GNOME modelling. In the near future, we will add new annotation tracks to the IGV genome browser, such as cell line-specific H3K27Ac marks, and colour these genomic features on 3D models to improve accessibility and facilitate better analysis of complex interactions between them.

DATA AVAILABILITY

3D-GNOME is freely available at <https://3dgnome.mini.pw.edu.pl/>.

ACKNOWLEDGEMENTS

We would like to thank Michał Dudeł for his assistance in resolving issues with the NGL viewer and Sebastian Korsak for his expertise regarding 3D model ensembles.

Computations were performed thanks to the Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology using the Artificial Intelligence HPC platform (Eden cluster) financed by Polish Ministry of Science and Higher Education (decision no. 7054/IA/SP/2020 of 2020-08-28).

FUNDING

Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme; Polish National Science Centre [2019/35/O/ST6/02484 and 2020/37/B/NZ2/03757]; EU-funded the Marie Skłodowska-Curie action (MSCA) Innovative Training Network named Enhpathy (www.enhpathy.eu) ‘Molecular Basis of Human enhanceropathies’. Funding for open

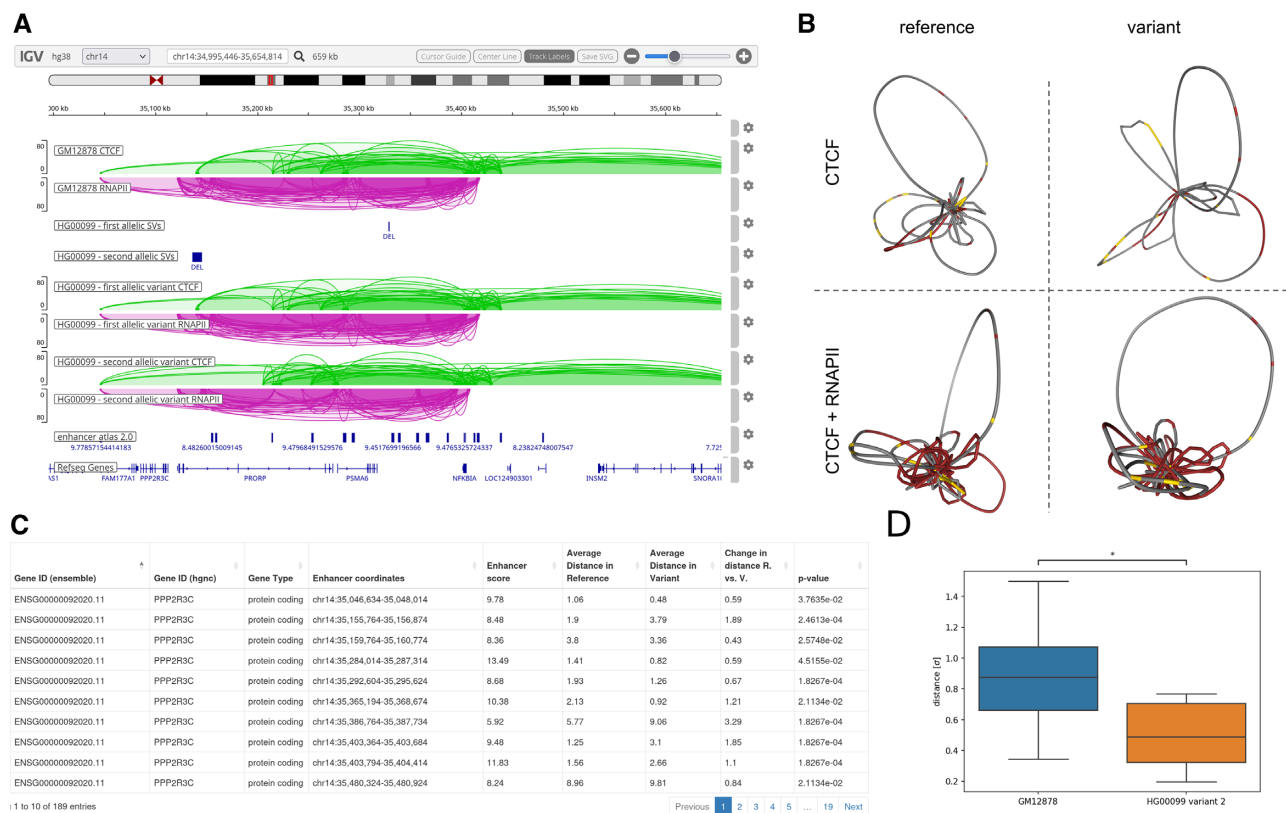


Figure 2. Results for the region chr14:35605439–35615196: (A) Screenshot of the IGV browser showing CTCF-mediated (green) and RNAPII-mediated (purple) chromatin interactions, as well as genomic annotations for the selected region. Tracks 1–2 display arcs for the reference cell line (GM12878), tracks 3–4 show SVs mapped to the reference genome, and tracks 5–8 show the arcs for variant I and variant II, respectively. Track 9 shows the genes located in the selected locus. (B) 3D chromatin models reconstructed for the selected region for the reference cell line (left) and variant II (right) based on CTCF interactions (up) and CTCF + RNAPII interactions (down). (C) A responsive table of the distances between the NFKB1 gene and enhancers in the reference and variant. (D) Box plots of the distance distribution between the NFKB1 gene and enhancer located in the chr14:35046634–35048014 region for the reference and variant 2.

access charge: Warsaw University of Technology, Warsaw, Poland.
Conflict of interest statement. None declared.

REFERENCES

1. Isbel, L., Grand, R.S. and Schübeler, D. (2022) Generating specificity in genome regulation through transcription factor sensitivity to chromatin. *Nat. Rev. Genet.*, **23**, 728–740.

2. Hafner, A. and Boettiger, A. (2022) The spatial organization of transcriptional control. *Nat. Rev. Genet.*, 1–16.

3. Chliński, M., Sengupta, K. and Plewczynski, D. (2022) From dna human sequence to the chromatin higher order organisation and its biological meaning: using biomolecular interaction networks to understand the influence of structural variation on spatial genome organisation and its functional effect. *Semin. Cell Dev. Biol.*, 171–185.

4. Farnham, P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.

5. Heintzman, N.D. and Ren, B. (2009) Finding distal regulatory elements in the human genome. *Curr. Opin. Genet. Dev.*, **19**, 541–549.

6. Levine, M. (2010) Transcriptional enhancers in animal development and evolution. *Curr. Biol.*, **20**, R754–R763.

7. Schoenfelder, S. and Fraser, P. (2019) Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.*, **20**, 437–455.

8. Scott, A.J., Hall, I.M. and Chiang, C. (2021) Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res.*, **526**, 2249–2257.

9. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsdóttir, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E. *et al.* (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.*, **95**, 535–552.

10. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory dna. *Science*, **337**, 1190–1195.

11. Liu, Y., Qu, H.-Q., Mentch, F.D., Qu, J., Chang, X., Nguyen, K., Tian, L., Glessner, J., Sleiman, P.M. and Hakonarson, H. (2022a) Application of deep learning algorithm on whole genome sequencing data uncovers structural variants associated with multiple mental disorders in african american patients. *Mol. Psychiatry*, **27**, 1469–1478.

12. Liu, Z., Roberts, R., Mercer, T.R., Xu, J., Sedlazeck, F.J. and Tong, W. (2022b) Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol.*, **23**, 68.

13. Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y. and Plewczynski, D. (2020) 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Res.*, **48**, W170–W176.

14. Szalaj, P., Michalski, P.J., Wróblewski, P., Tang, Z., Kadlof, M., Mazzocco, G., Ruan, Y. and Plewczynski, D. (2016) 3D-GNOME: an integrated web service for structural modeling of the 3D genome. *Nucleic Acids Res.*, **44**, W288–W293.

15. Consortium, T.I.G.P. (2015) A global reference for human genetic variation. *J. Open Source Software*, **526**, 68–74.

16. Szalaj, P., Tang, Z., Michalski, P., Pietal, M.J., Luo, O.J., Sadowski, M., Li, X., Radew, K., Ruan, Y. and Plewczynski, D. (2016) An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization. *Genome Res.*, **26**, 1697–1709.

17. Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinf.*, **14**, 178–192.
18. Rose, A.S. and Hildebrand, P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
19. Kadlof, M., Rozycka, J. and Plewczynski, D. (2020) Spring Model—chromatin modeling tool based on OpenMM. *Methods*, **181**, 62–69.
20. Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M.N.K., Li, Y., Hu, M. *et al.* (2018) The 3D genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.*, **19**, 1–12.
21. Li, X., Luo, O.J., Wang, P., Zheng, M., Wang, D., Piecuch, E., Zhu, J.J., Tian, S.Z., Tang, Z., Li, G. *et al.* (2017) Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nat. Protoc.*, **12**, 899–915.
22. Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Włodarczyk, J., Ruszczyski, B. *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
23. Wang, P., Feng, Y., Zhu, K., Chai, H., Chang, Y., Yang, X., Liu, X., Shen, C., Gega, E., Lee, B. *et al.* (2021) In situ chromatin interaction analysis using paired-end tag sequencing. *Curr. Protoc.*, **1**, e174.
24. Yoo, A.B., Jette, M.A. and Grondona, M. (2003) Slurm: simple linux utility for resource management. In: *Job Scheduling Strategies for Parallel Processing: 9th international Workshop, JSSPP 2003, Seattle, WA, USA, June 24, 2003. Revised Paper 9*. Springer. pp.44–60.
25. Tange, O. (2011) Gnu parallel—the command-line power tool. *USENIX Mag.*, **36**, 42–47.
26. Gao, T. and Qian, J. (2019) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.*, **48**, D58–D64.
27. Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y. and Plewczynski, D. (2019) Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome Biol.*, **20**, 148.
28. Waskom, M.L. (2021) Seaborn: statistical data visualization. *J. Open Source Softw.*, **6**, 3021.

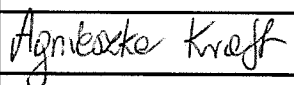
Declaration

I hereby declare that the contribution to the following paper:

Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y., & Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome biology*, 20(1), 1-27. DOI: <https://doi.org/10.1186/s13059-019-1728-x>, URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1728-x>

is correctly characterized below:

DP and MS conceived and implemented the methodology of modeling CCDs of individual genomes at the population scale. PS, DP, ZT, and YR devised the 3D-GNOME used as the main engine for modeling. MW extended the 3D-GNOME web service to provide the SV-including modeling method (3D-GNOME 2.0). MS designed the statistical analysis part. MS and AK performed the analyses. MS, AK, and ZT extracted and prepared the data for the analyses. MS, PS, AK, ZT, YR, and DP prepared the manuscript. MS was a major contributor in writing the manuscript. DP, YR, and ZT supervised the study. All authors read and approved the final manuscript.

Name	Abbreviation	Signature
Michał Sadowski	MS	
Agnieszka Kraft	AK	
Przemysław Szalaj	PS	
Michał Wlasnowolski	MW	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	

June 01, 2023

Declaration

I hereby declare that the contribution to the following paper:

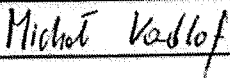
Wlasnowolski, M., Kadlof, M., Sengupta, K., & Plewczynski, D. (2023). 3D-GNOME 3.0: a three-dimensional genome modelling engine for analysing changes of promoter-enhancer contacts in the human genome. *Nucleic Acids Research*, gkad354.

DOI: <https://doi.org/10.1093/nar/gkad354>

URL: <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad354/7157515>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: integration of new datasets and data formats with the web server, development of the web server update. MK and KS: preparation of the figures. MK and MW: integration of the web service with the Eden cluster. MW, MK and KS: contribution to writing the text, with MW being a major contributor in writing the manuscript. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Wlasnowolski	MW	
Michał Kadlof	MK	
Kaustav Sengupta	KS	
Dariusz Plewczynski	DP	


Declaration

I hereby declare that the contribution to the following paper:

Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y. and Plewczynski, D., 2020. 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Research*, [online] 48(W1), pp.W170–W176.
DOI: <https://doi.org/10.1093/nar/gkaa388>
URL: <https://academic.oup.com/nar/article/48/W1/W170/5842186>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: integration of new datasets and data formats with the web server, and developed the web server update. MW and KJ figure preparation. MW, MS, and KJ text writing. TC feature development for labelling structural variants on arc diagrams. PS data browser development. ZT and YR preparation of the default CTCF and RNAPII interaction dataset. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Michał Sadowski	MS	
Tymon Czarnota	TC	
Karolina Jodkowska	KJ	
Przemysław Szalaj	PS	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	

Declaration

I hereby declare that the contribution to the following paper:

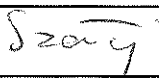
Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y. and Plewczynski, D., 2020. 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Research*, [online] 48(W1), pp.W170–W176.

DOI: <https://doi.org/10.1093/nar/gkaa388>

URL: <https://academic.oup.com/nar/article/48/W1/W170/5842186>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: integration of new datasets and data formats with the web server, and developed the web server update. MW and KJ figure preparation. MW, MS, and KJ text writing. TC feature development for labelling structural variants on arc diagrams. PS data browser development. ZT and YR preparation of the default CTCF and RNAPII interaction dataset. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Wlasnowolski	MW	
Michał Sadowski	MS	
Tymon Czarnota	TC	
Karolina Jodkowska	KJ	
Przemysław Szalaj	PS	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	

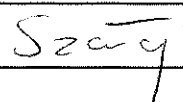

Declaration

I hereby declare that the contribution to the following paper:

Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y., & Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome biology*, 20(1), 1-27. DOI: <https://doi.org/10.1186/s13059-019-1728-x>, URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1728-x>

is correctly characterized below:

DP and MS conceived and implemented the methodology of modeling CCDs of individual genomes at the population scale. PS, DP, ZT, and YR devised the 3D-GNOME used as the main engine for modeling. MW extended the 3D-GNOME web service to provide the SV-including modeling method (3D-GNOME 2.0). MS designed the statistical analysis part. MS and AK performed the analyses. MS, AK, and ZT extracted and prepared the data for the analyses. MS, PS, AK, ZT, YR, and DP prepared the manuscript. MS was a major contributor in writing the manuscript. DP, YR, and ZT supervised the study. All authors read and approved the final manuscript.

Name	Abbreviation	Signature
Michał Sadowski	MS	
Agnieszka Kraft	AK	
Przemysław Szalaj	PS	
Michał Własnowolski	MW	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	


Declaration

I hereby declare that the contribution to the following paper:

Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y., & Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome biology*, 20(1), 1-27. DOI: <https://doi.org/10.1186/s13059-019-1728-x>, URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1728-x>

is correctly characterized below:

DP and MS conceived and implemented the methodology of modeling CCDs of individual genomes at the population scale. PS, DP, ZT, and YR devised the 3D-GNOME used as the main engine for modeling. MW extended the 3D-GNOME web service to provide the SV-including modeling method (3D-GNOME 2.0). MS designed the statistical analysis part. MS and AK performed the analyses. MS, AK, and ZT extracted and prepared the data for the analyses. MS, PS, AK, ZT, YR, and DP prepared the manuscript. MS was a major contributor in writing the manuscript. DP, YR, and ZT supervised the study. All authors read and approved the final manuscript.

Name	Abbreviation	Signature
Michał Sadowski	MS	
Agnieszka Kraft	AK	
Przemysław Szalaj	PS	
Michał Własnowolski	MW	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	

Declaration

I hereby declare that the contribution to the following paper:


Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y. and Plewczynski, D., 2020. 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Research*, [online] 48(W1), pp.W170–W176.

DOI: <https://doi.org/10.1093/nar/gkaa388>

URL: <https://academic.oup.com/nar/article/48/W1/W170/5842186>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: integration of new datasets and data formats with the web server, and developed the web server update. MW and KJ figure preparation. MW, MS, and KJ text writing. TC feature development for labelling structural variants on arc diagrams. PS data browser development. ZT and YR preparation of the default CTCF and RNAPII interaction dataset. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Michał Sadowski	MS	
Tymon Czarnota	TC	
Karolina Jodkowska	KJ	
Przemysław Szalaj	PS	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	


Declaration

I hereby declare that the contribution to the following paper:

Własnowolski, M., Grabowski, P., Roszczyk, D., Kaczmarek, K., & Plewczynski, D.
(2023). cudaMMC - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling.
bioRxiv. <https://doi.org/10.1101/2023.06.12.544609>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: method testing and results validation, major contributor in writing the manuscript, figure preparation (graphs and 3D model). PG and DR implementation of the scattering of Monte Carlo simulated annealing computations on GPU cards, pseudocode and figure preparation (activity diagram), KK: supervision of the GPU-related tasks and writing contribution to the GPU-related sections of the manuscript. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Paweł Grabowski	PG	
Damian Roszczyk	DR	
Krzysztof Kaczmarek	KK	
Dariusz Plewczyński	DP	

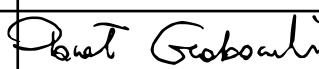
Declaration

I hereby declare that the contribution to the following paper:

Wlasnowolski, M., Grabowski, P., Roszczyk, D., Kaczmariski, K., & Plewczynski, D.
(2023). cudaMMC - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling.
bioRxiv. <https://doi.org/10.1101/2023.06.12.544609>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: method testing and results validation, major contributor in writing the manuscript, figure preparation (graphs and 3D model). PG and DR implementation of the scattering of Monte Carlo simulated annealing computations on GPU cards, pseudocode and figure preparation (activity diagram), KK: supervision of the GPU-related tasks and writing contribution to the GPU-related sections of the manuscript. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Paweł Grabowski	PG	
Damian Roszczyk	DR	
Krzysztof Kaczmariski	KK	
Dariusz Plewczyński	DP	


Declaration

I hereby declare that the contribution to the following paper:

Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y., & Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome biology*, 20(1), 1-27. DOI: <https://doi.org/10.1186/s13059-019-1728-x>, URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1728-x>

is correctly characterized below:

DP and MS conceived and implemented the methodology of modeling CCDs of individual genomes at the population scale. PS, DP, ZT, and YR devised the 3D-GNOME used as the main engine for modeling. MW extended the 3D-GNOME web service to provide the SV-including modeling method (3D-GNOME 2.0). MS designed the statistical analysis part. MS and AK performed the analyses. MS, AK, and ZT extracted and prepared the data for the analyses. MS, PS, AK, ZT, YR, and DP prepared the manuscript. MS was a major contributor in writing the manuscript. DP, YR, and ZT supervised the study. All authors read and approved the final manuscript.

Name	Abbreviation	Signature
Michał Sadowski	MS	
Agnieszka Kraft	AK	
Przemysław Szalaj	PS	
Michał Własnowolski	MW	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	

Declaration

I hereby declare that the contribution to the following paper:


Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y. and Plewczynski, D., 2020. 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Research*, [online] 48(W1), pp.W170–W176.

DOI: <https://doi.org/10.1093/nar/gkaa388>

URL: <https://academic.oup.com/nar/article/48/W1/W170/5842186>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: integration of new datasets and data formats with the web server, and developed the web server update. MW and KJ figure preparation. MW, MS, and KJ text writing. TC feature development for labelling structural variants on arc diagrams. PS data browser development. ZT and YR preparation of the default CTCF and RNAPII interaction dataset. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Michał Sadowski	MS	
Tymon Czarnota	TC	
Karolina Jodkowska	KJ	
Przemysław Szalaj	PS	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	


Declaration

I hereby declare that the contribution to the following paper:

Wlasnowolski, M., Grabowski, P., Roszczyk, D., Kaczmarek, K., & Plewczynski, D.
(2023). cudaMMC - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling.
bioRxiv. <https://doi.org/10.1101/2023.06.12.544609>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: method testing and results validation, major contributor in writing the manuscript, figure preparation (graphs and 3D model). PG and DR implementation of the scattering of Monte Carlo simulated annealing computations on GPU cards, pseudocode and figure preparation (activity diagram), KK: supervision of the GPU-related tasks and writing contribution to the GPU-related sections of the manuscript. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Paweł Grabowski	PG	
Damian Roszczyk	DR	
Krzysztof Kaczmarek	KK	
Dariusz Plewczyński	DP	

June 01, 2023

Declaration

I hereby declare that the contribution to the following paper:


Wlasnowolski, M., Kadlof, M., Sengupta, K., & Plewczynski, D. (2023). 3D-GNOME 3.0: a three-dimensional genome modelling engine for analysing changes of promoter-enhancer contacts in the human genome. *Nucleic Acids Research*, gkad354.

DOI: <https://doi.org/10.1093/nar/gkad354>

URL: <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad354/7157515>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: integration of new datasets and data formats with the web server, development of the web server update. MK and KS: preparation of the figures. MK and MW: integration of the web service with the Eden cluster. MW, MK and KS: contribution to writing the text, with MW being a major contributor in writing the manuscript. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Michał Kadlof	MK	
Kaustav Sengupta	KS	
Dariusz Plewczynski	DP	

June 01, 2023

Declaration

I hereby declare that the contribution to the following paper:


Wlasnowolski, M., Kadlof, M., Sengupta, K., & Plewczynski, D. (2023). 3D-GNOME 3.0: a three-dimensional genome modelling engine for analysing changes of promoter-enhancer contacts in the human genome. *Nucleic Acids Research*, gkad354.

DOI: <https://doi.org/10.1093/nar/gkad354>

URL: <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad354/7157515>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: integration of new datasets and data formats with the web server, development of the web server update. MK and KS: preparation of the figures. MK and MW: integration of the web service with the Eden cluster. MW, MK and KS: contribution to writing the text, with MW being a major contributor in writing the manuscript. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Michał Kadlof	MK	
Kaustav Sengupta	KS	
Dariusz Plewczynski	DP	

June 01, 2023

Declaration

I hereby declare that the contribution to the following paper:

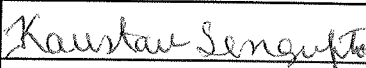
Wlasnowolski, M., Kadlof, M., Sengupta, K., & Plewczynski, D. (2023). 3D-GNOME 3.0: a three-dimensional genome modelling engine for analysing changes of promoter-enhancer contacts in the human genome. *Nucleic Acids Research*, gkad354.

DOI: <https://doi.org/10.1093/nar/gkad354>

URL: <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad354/7157515>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: integration of new datasets and data formats with the web server, development of the web server update. MK and KS: preparation of the figures. MK and MW: integration of the web service with the Eden cluster. MW, MK and KS: contribution to writing the text, with MW being a major contributor in writing the manuscript. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Michał Kadlof	MK	
Kaustav Sengupta	KS	
Dariusz Plewczynski	DP	

June 13, 2023

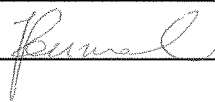
Declaration

I hereby declare that the contribution to the following paper:

Własnowolski, M., Grabowski, P., Roszczyk, D., Kaczmarowski, K., & Plewczynski, D.
(2023). cudaMMC - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling.
bioRxiv. <https://doi.org/10.1101/2023.06.12.544609>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: method testing and results validation, major contributor in writing the manuscript, figure preparation (graphs and 3D model). PG and DR implementation of the scattering of Monte Carlo simulated annealing computations on GPU cards, pseudocode and figure preparation (activity diagram), KK: supervision of the GPU-related tasks and writing contribution to the GPU-related sections of the manuscript. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Paweł Grabowski	PG	
Damian Roszczyk	DR	
Krzysztof Kaczmarowski	KK	
Dariusz Plewczyński	DP	

June 13, 2023


Declaration

I hereby declare that the contribution to the following paper:

Wlasnowolski, M., Grabowski, P., Roszczyk, D., Kaczmarek, K., & Plewczynski, D.
(2023). cudaMMC - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling.
bioRxiv. <https://doi.org/10.1101/2023.06.12.544609>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: method testing and results validation, major contributor in writing the manuscript, figure preparation (graphs and 3D model). PG and DR implementation of the scattering of Monte Carlo simulated annealing computations on GPU cards, pseudocode and figure preparation (activity diagram), KK: supervision of the GPU-related tasks and writing contribution to the GPU-related sections of the manuscript. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Paweł Grabowski	PG	
Damian Roszczyk	DR	
Krzysztof Kaczmarek	KK	
Dariusz Plewczyński	DP	

June 01, 2023

Declaration

I hereby declare that the contribution to the following paper:


Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y. and Plewczynski, D., 2020. 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Research*, [online] 48(W1), pp.W170–W176.

DOI: <https://doi.org/10.1093/nar/gkaa388>

URL: <https://academic.oup.com/nar/article/48/W1/W170/5842186>

is correctly characterized below:

MW and DP: team coordination, conceptualization, and methodology. MW: integration of new datasets and data formats with the web server, and developed the web server update. MW and KJ figure preparation. MW, MS, and KJ text writing. TC feature development for labelling structural variants on arc diagrams. PS data browser development. ZT and YR preparation of the default CTCF and RNAPII interaction dataset. DP: funding acquisition and supervision of the study.

Name	Abbreviation	Signature
Michał Własnowolski	MW	
Michał Sadowski	MS	
Tymon Czarnota	TC	
Karolina Jodkowska	KJ	
Przemysław Szalaj	PS	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	

June 01, 2023

Declaration

I hereby declare that the contribution to the following paper:


Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y., & Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome biology*, 20(1), 1-27. DOI:

<https://doi.org/10.1186/s13059-019-1728-x>, URL:

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1728-x>

is correctly characterized below:

DP and MS conceived and implemented the methodology of modeling CCDs of individual genomes at the population scale. PS, DP, ZT, and YR devised the 3D-GNOME used as the main engine for modeling. MW extended the 3D-GNOME web service to provide the SV-including modeling method (3D-GNOME 2.0). MS designed the statistical analysis part. MS and AK performed the analyses. MS, AK, and ZT extracted and prepared the data for the analyses. MS, PS, AK, ZT, YR, and DP prepared the manuscript. MS was a major contributor in writing the manuscript. DP, YR, and ZT supervised the study. All authors read and approved the final manuscript.

Name	Abbreviation	Signature
Michał Sadowski	MS	
Agnieszka Kraft	AK	
Przemysław Szalaj	PS	
Michał Własnowolski	MW	
Zhonghui Tang	ZT	
Yijun Ruan	YR	
Dariusz Plewczyński	DP	

Copies of additional publications not included in the collection:

Herman-Izycka, J., **Wlasnowolski, M.**, & Wilczynski, B. (2017). Taking promoters out of enhancers in sequence based predictions of tissue-specific mammalian enhancers. *BMC medical genomics*, 10, 17-26.

IF=3.622, MNiSW points: 100

Contribution: analyzed predictions in context of DHS sites and created website providing predictions.

Sarkar, J. P., Saha, I., Rakshit, S., Pal, M., **Wlasnowolski, M.**, Sarkar, A., Maulik, U., & Plewczynski, D. (2019, July). A new evolutionary rough fuzzy integrated machine learning technique for microRNA selection using next-generation sequencing data of breast cancer. In Proceedings of the *Genetic and Evolutionary Computation Conference Companion* (pp. 1846-1854).

MNiSW points: 140, related to ITT discipline

Contribution: results validation.

Sarkar, J. P., Saha, I., Lancucki, A., Ghosh, N., **Wlasnowolski, M.**, Bokota, G., Dey, A., Lipinski, P., & Plewczynski, D. (2020). Identification of miRNA biomarkers for diverse cancer types using statistical learning methods at the whole-genome scale. *Frontiers in Genetics*, 11, 982.

IF=4.772, MNiSW points: 100

Contribution: co-writing the manuscript.

Saha, I., Rakshit, S., **Wlasnowolski, M.**, & Plewczynski, D. (2019, October). Identification of epigenetic biomarkers with the use of gene expression and DNA methylation for breast cancer subtypes. In *Tencon 2019-2019 Ieee Region 10 Conference (Tencon)* (pp. 417-422). IEEE.

MNiSW points: 20, related to ITT discipline

Contribution: data preprocessing, manuscript writing.

RESEARCH

Open Access



Taking promoters out of enhancers in sequence based predictions of tissue-specific mammalian enhancers

Julia Herman-Izycka, Michal Wlasnowolski and Bartek Wilczynski*

From The 6th Translational Bioinformatics Conference
Je Ju Island, Korea. 15–17 October 2016

Abstract

Background: Many genetic diseases are caused by mutations in non-coding regions of the genome. These mutations are frequently found in enhancer sequences, causing disruption to the regulatory program of the cell. Enhancers are short regulatory sequences in the non-coding part of the genome that are essential for the proper regulation of transcription. While the experimental methods for identification of such sequences are improving every year, our understanding of the rules behind the enhancer activity has not progressed much in the last decade. This is especially true in case of tissue-specific enhancers, where there are clear problems in predicting specificity of enhancer activity.

Results: We show a random-forest based machine learning approach capable of matching the performance of the current state-of-the-art methods for enhancer prediction. Then we show that it is, similarly to other published methods, frequently cross-predicting enhancers as active in different tissues, making it less useful for predicting tissue specific activity. Then we proceed to show that the problem is related to the fact that the enhancer predicting models exhibit a bias towards predicting gene promoters as active enhancers. Then we show that using a two-step classifier can lead to lower cross-prediction between tissues.

Conclusions: We provide whole-genome predictions of human heart and brain enhancers obtained with two-step classifier.

Keywords: Enhancer prediction, Regulatory sequence, Histone modifications, Machine learning

Background

Transcription regulation is a complex process requiring tight control at multiple steps including transcription initiation, elongation and splicing. In case of tissue specific genes in metazoan genomes, the control of the transcription initiation is performed largely by means of enhancers, i.e. distinct sequence elements, that allow for binding of transcription factor proteins, facilitating transcription [1]. While the exact molecular mechanism of enhancer-promoter interaction remains a field of active study, we have now accumulated a large body of examples

of enhancer sequences to ask whether we can make predictions regarding enhancer location in the genome based on sequence features. This is an important question, given the complexity of a gene regulation system in a multicellular organism such as humans composed of hundreds of cell types, each of which expresses thousands of genes, most of which are modulated with some cell-type specificity. Moreover, a typical cell-type-specific gene can be controlled by multiple enhancers. Adding it all up, in order to describe a tissue-specific gene regulation, we need to describe on the order of 100 thousands of regulator elements [2].

Mapping all these elements using experimental techniques is currently completely unfeasible, as many cell-types are too difficult to obtain in large quantities required

*Correspondence: bartek@mimuw.edu.pl
Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland

for experimental assessment of enhancer activity. This leads to a situation, where we have hundreds of well documented examples of regulatory elements functional in a certain context (i.e. cell-type, developmental time) determined by a certain method (enhancer reporter assays [3], STARR-Seq [4], luciferase assays, in-situ hybridization etc). This, however, cannot easily be scaled up to the level of complete coverage of all cell types and all developmental time points. On the other hand, the data collected by the ENCODE or Epigenome Roadmap [5] are invaluable as a source for computational attempts at making models that would be predictive beyond the collected data and perhaps eventually help defining the principles of tissue-specific action of regulatory elements.

Even before the complete sequence of the human genome was known, people have been working on computational descriptions of enhancer sequences – mostly based on clusters of transcription factor binding sites [6, 7]. Later, these models were improved by including evolutionary conservation of sequence features [8] leading to a classical now approach of predicting enhancers as evolutionarily conserved clusters of transcription factor binding sites. While a number of methods following this theme, but varying greatly on the technical side [9–11] has been proposed in the first decade of 2000s, their relative performance was inherently limited by the imperfect training data [12, 13]. The limitations were not only due to the small training sets but also the difficulty to assess the true quality of predictions.

Importantly, while the databases containing the position weight matrices describing transcription factor binding specificity grew rapidly, it became clear that they are not necessarily the best representations of the important sequence features for enhancer predictions [14]. This is due to two main reasons: large similarity of many transcription factor's binding domains, leading to very similar DNA specificity and the frequently artificial specificity encoded in the position weight matrices based on context-specific determination of binding. Taking the two together, the sequence motif databases were not optimal for the task and it has been shown that the same or better accuracy in enhancer prediction can be achieved with counting k-mers, instead of the actual transcription factor motifs [15].

The advent of the ChIP-Seq technology [16], allowing researchers to directly assay transcription factor binding as well as multiple histone modifications, changed the situation. The availability of genome-wide measurements of transcription factor binding enabled much more comprehensive training of the predictors based on generic machine learning methods [17, 18], however they uncovered an unanticipated complexity of enhancer activity. In particular, the ChIP-Seq based methods allowed us to uncover many regulatory elements that were clearly

functional without a detectable sequence conservation between species [17], and the studies of transcription factor binding across developmental time-points or conditions detected large scale context dependency of enhancer function [19]. These findings were later corroborated by multiple studies using DNase-Seq methods [20, 21].

The wide adoption of ChIP-Seq technique together with concentrated experimental efforts of ENCODE and similar allowed for the new wave of computational approaches to appear. Typically these would take advantage of hundreds genome-wide tracks of ChIP-Seq, DNase-Seq and other information (such as mRNA-Seq or GRO-Seq) and use a state-of-the-art machine learning method such as SVM [22, 23], Bayesian Networks [18], random forests [24, 25], neural networks [26] or regression based, such as support vector regression [27] or logistic regression [28]. The first study by Erwin et al. should be noted especially for their careful analysis of the specificity of the classifiers. They observed, that when they trained their models on positive vs. random data sets, the resulting classifiers gave overlapping predictions between tissues. Erwin et al. used a specific way of two-layer classification, where they first predicted whether a sequence is an enhancer and in the second step they actually predicted to which class from the training set it belongs. This is an important observation, however, the solution leaves room for improvement as it yields classifiers that are only capable of discerning between the activity classes known at the time of learning. In particular, using this approach one cannot make any claims regarding the overlap of predictions with the regions with activity in any other tissue.

It should be noted that many of these studies differ significantly also in their choice of the training sets. For example Zhu et al. [28] use eRNAs as their positive set, Danko et al. [27] utilize GRO-Seq data, Firpi et al. [26] use histone modification ChIP signals preprocessed by Heintzmann et al. [17] while Rajagopal [25] uses p300 distal binding sites. All of these are then very difficult to compare with approaches such as ours or Erwin et al's [22] that use Vista enhancers.

While enhancer predictions in these studies have reached the level of approximately 90 percent of AUC (Area under the ROC curve) in cross-validated setup, their ability to help us understand the function of enhancers is limited, partially because of their foundation on a very large set of measurements and somewhat opaque machine learning approaches. This has prompted us to approach this problem with a slightly different methodology. We have focused on the issue of enhancer tissue-specificity, i.e. the ability of enhancer sequences to be active only under a very defined set of circumstances and to define a set of features that are crucial for predicting the activity. This has led us to focus less on quality of predictions, but still above “acceptable” level of 80

percent AUC, while retaining the possibility of rigorously assessing the relative importance of the features used for prediction. This has lead to many interesting predictions that are very specific to heart or brain (See Additional file 1: Figure S8, S9 and S10 for representative examples).

In this work we report our findings based on applying Random Forest classification to the problem of enhancer prediction in the human genome. Based on our previous experiences with the *Drosophila* [24, 29], we have used histone modification ChIP-Seq datasets from ENCODE and the enhancer sequences from the Vista project. Using these data, we have defined a set of features that are relevant to define active enhancers and then we went to assess the tissue specificity of the prediction using the heart and brain tissue annotation from the Vista project. This has lead to our discovery that both groups share an enrichment for predictions in the group of proximal-promoter sequences leading to two kinds of problems: firstly both classifiers predict promoters as enhancer and secondly, both classifiers use shared features to predict promoter regions leading to further overlap in genome-wide predictions. In order to tackle this problem we proposed building two-layer classifier, where one layer is responsible for detecting tissue-specific enhancer signals while the other is a pure sequence-based filter of promoter-like sequences that ensures greater specificity of the complete classifier.

Methods

Training set and data preparation

Positive training set We downloaded all human sequences available in VISTA Enhancer Browser on April 15th, 2014. Our heart training set consists of sequences that show heart activity (among others) but no brain activity, and brain training set is defined as sequences with activity in at least one of those tissues: hindbrain, midbrain, forebrain, neural tube, cranial nerve, but does not show heart activity. Non-specific classifier is trained on both heart and brain sets, defined as above.

Negative training set Negative training sets were chosen randomly from the human genome hg19 in the way that they preserve chromosome and length distribution the same as in corresponding positive set. We draw lengths of sequences from Negative Binomial Distribution with parameters matching mean and variance of lengths of sequences in the positive training set. Since heart enhancers in VISTA database are on average longer than brain enhancers (see Additional file 1: Figure S3), we draw separate negative training sets for heart and brain classifiers. We ensure we do not include sequences with ambiguous bases and sequences that have less than 10% of signal in more than quarter of considered histone modifications tracks.

Sequence features Our sequence feature set consist of all possible k -mers – continuous sequences of k bases, excluding second alphabetically k -mer for each pair of reverse complement k -mers. Each feature is represented by number of occurrences of k -mer (or its reverse complement) over length of sequence. Since our random set is chosen in the way, that it doesn't contain ambiguous bases, we have almost no such bases in the sequences so while counting we skip k -mers containing 'N'.

Histone modifications features Histone modification signal was obtained from ChIP-Seq experiment results from ENCODE database. We used downloaded normalized signal for all histone modifications available for cells from Tier1 (H1hESC, GM12878, K562) and normal (non-cancer) cells from Tier2 (CD20+, CD20+_RO01778, CD20+_RO01794, HUVEC, Monocytes-CD14+_RO01746). For list of files see Additional file 1: Table S1. Each ChIP-Seq track contributes one feature to our dataset – mean signal over considered region (e.g. enhancer).

Training and performance of classifiers

We used Python implementation of random forest from scikit-learn version 0.14.1. We trained initial classifiers with 100 and 1000 trees. Results in terms of AUC were very similar (improvement by 0.02 to 0.16 points between 100 and 1000 trees), so in favor of time we decided to use 100-trees classifier. We used Gini criterion for split quality (default), as well as other default options, which for example ensures building trees that split training samples perfectly.

Each training was performed with positive and negative training sets of equal sizes. If necessary, when positive and negative sets had different size, subset of samples is drawn from larger set before training.

For performance assessment we used stratified 10-fold cross-validation and computed Area Under Receiver-Operator Characteristic Curve (AUC). Presented values are mean AUC from 10 rounds of training (with, when necessary, independently drawn subsets of our training sets).

P -value of difference of AUC between two classifiers trained on same training set (but different features) is computed as presented in [30], i.e. it is the p -value (two tailed) for a z -score of tested AUC against expected AUC.

Feature importance

We run Boruta algorithm ([31], version 3.1), a wrapper around Random Forest, that allows assessment of feature importance comparing to random features. It runs multiple rounds of classification (here up to 100) and compares importance of each feature (defined as Z -score of loss of accuracy if feature is permuted), with importance of

random features (shuffled initial features). Features that where significantly (p -value 0.01, Bonferroni correction), more often more important than best random feature are marked as Important, more often less important than best random feature are marked as Unimportant, and after all of rounds the rest is marked as Tentative.

Whole-genome predictions

To compute whole-genome predictions we divided hg19 genome (all autosomal chromosomes) into windows of length 1500 bp, every 750 bp. We excluded windows with ambiguous sequence (i.e. containing at least one 'N'). For each window we output probability of being enhancer – a result of Random Forest voting (as fraction of trees that predict a sequence is active).

Comparison with DNase hypersensitivity sites

To compare enhancers predictions with DNase hypersensitivity sites we downloaded DNase clusters data (V3) from ENCODE database (<https://genome.ucsc.edu/ENCODE/>) and DNase ChIP-Seq data from Roadmap Epigenomics database (<http://www.roadmapepigenomics.org/>). We compare non-specific 4-mers classifier's predictions on DHS windows (overlapping DHS clusters by at least 100 bp) with non-DHS windows (other). We excluded windows containing TSS and plot results for 1000 randomly selected DHS and non-DHS windows from the 1-st chromosome.

To define tissue-specific DHS and non-DHS windows we use DNase ChIP-Seq signal. One thousand windows with highest aggregate signal create DHS set, non-DHS set of 1000 windows is drawn randomly from all windows with maximal signal smaller than 10. We compared distribution of prediction for those sets.

Promoter predictions and two-step classifiers

We define promoter predictions as those windows from whole-genome 4-mers predictions that have high score (> 0.8) and contain at least one TSS from the list of 215,881 TSS from ENSEMBL (downloaded on July 1, 2015). Classifiers trained on promoter predictions are trained on 50 (out of 1775) heart and 50 (out of 632) randomly chosen brain predicted promoters. Second-step classifiers trained on random or VISTA sequences use sequences with length adjusted (extended or shrunk) to 1500 bp, but maintaining same middle position.

Validation of predictions on new VISTA sequences

After initial training of the classifiers on VISTA sequences few more records were deposited to this database (5 heart and 3 brain enhancers, as of October 10th, 2016). We took the opportunity to validate our classifier on those sequences (see Additional file 1: Table S7). While heart 4-mers classifier predictions were distinguishing heart

from brain enhancers (average predictions 0.64 and 0.49), and two-step heart classifier rated heart sequences only slightly above brain (0.3 and 0.27), brain one-step classifier distinguished brain from heart sequences worse (avg. 0.78 and 0.71) than two-step approach (0.52 to 0.33).

Results

Predicting mammalian enhancers using random forest classifier

Our goal was to use supervised machine learning approach to build a method, that given a set of active enhancers and set of non-enhancers can predict probability of a sequence being active as an enhancer in same tissue as sequences from positive training set. We chose to use Random Forest classifier [32], which performs well on both small and larger feature sets, and enables assessment of importance of individual features. In this method many decision trees trained on subsets of training data are incorporated in order to get a more robust classifier. We trained classifiers on training datasets with various tissue-specificity, aiming to obtain tissue-specific enhancer classifiers (Fig. 1). We also used different feature sets to compare importance of various groups of features for prediction of enhancers.

As our training set we use experimentally validated active enhancers (as positive examples) and random genomic sequences (as a negative set). We took enhancers from VISTA Enhancer Browser [33], database that contains over thousand of human and mouse sequences tested in transgenic mice, with activity confirmed on particular moment of embryonic development. VISTA sequences are also annotated with tissue (or tissues) of activity. For tissue-specific prediction we chose heart and brain tissue, since these were the most abundant in our active enhancers database, and for non-specific we use both heart and brain active sequences.

Adding histone modifications data to sequence information improves prediction

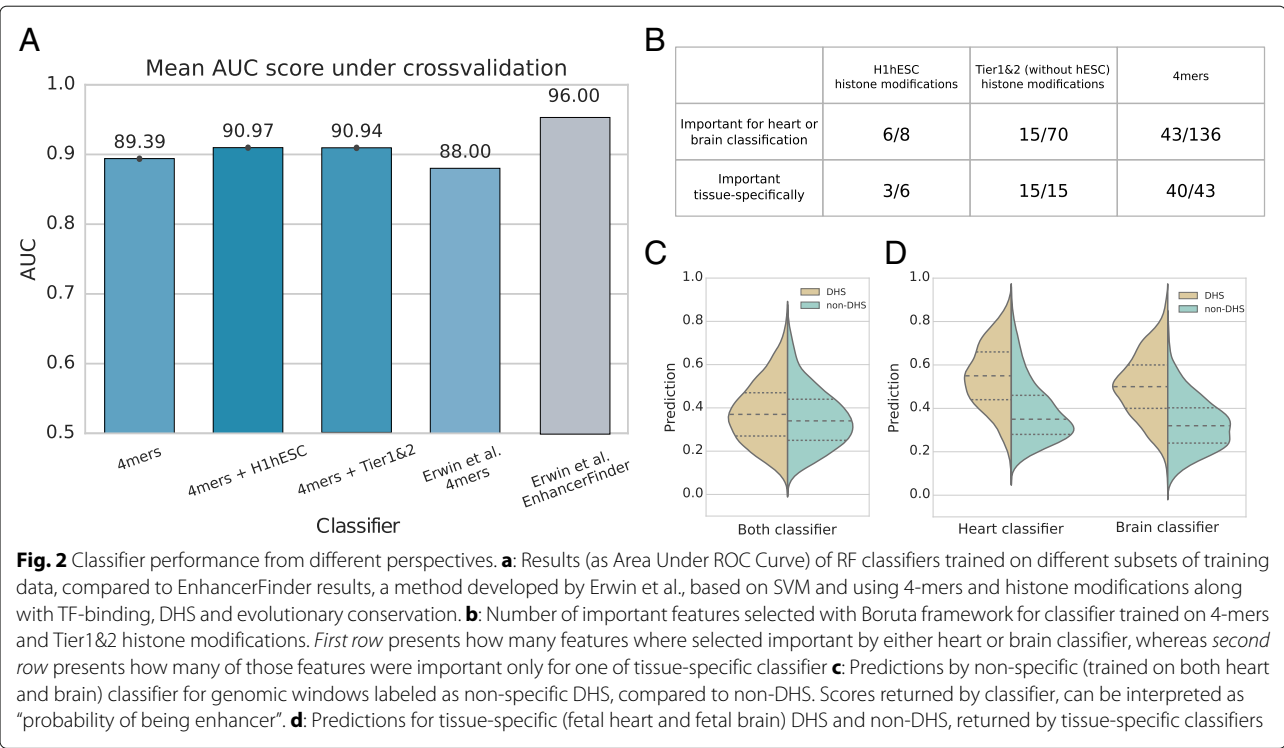
One of the goals of our work was to check whether random forest classifiers can be effective in combining two types of biological data: DNA sequence of a region and histone modifications within that region to improve prediction accuracy. SVM classifiers were previously shown to do so [22]. As a representation of sequence information we use frequency of k -mers — words of length k . k -mers are simpler model than TFBS motifs and they do not limit information representation only to sequences that represent known TFBS motifs. We took $k = 4$, as it gives us relatively small number of features (below 200) while giving slightly better results than 3-mers (see Table 1).

Histone modifications data came from ENCODE project ([34]). We used all available at the time ChIP-Seq tracks for histone modifications in the most well studied



Tissue	Kmers	Hmods	AUC
Non-specific	3 mers	–	0.871
Non-specific	4 mers	–	0.894
Non-specific	4 mers	H1hESC	0.910
Non-specific	4 mers	Tier1&2	0.910
Heart	3 mers	–	0.771
Heart	4mers	–	0.780
Heart	4 mers	H1hESC	0.809
Heart	4 mers	Tier1&2	0.844
Brain	3 mers	–	0.878
Brain	4 mers	–	0.906
Brain	4 mers	H1hESC	0.923
Brain	4 mers	Tier1&2	0.923

We performed multiple runs of classifier training using different subsets of our feature set and we compared obtained classifiers by measuring their Area Under ROC Curve (AUC). In every case Random Forest classifier gave results comparable with other state of the art methods (AUC between 0.87 and 0.93 — see Table 1). Addition of histone modification data from H1hESC or Tier1&2 improved prediction, although the changes are not significant (p -value $\cong 0.09$, computed as in [30]). More results can be found in the Table 1. These results are strikingly similar to the earlier results by Erwin et al. (see Fig. 2a)



especially when one considers the difference in size of feature sets (besides *k*-mers and evolutionary conservation 2496 features were used for EnhancerFinder training, comparing to 8 for our H1hESC or 78 for Tier1&2 classifier). While the best performance of EnhancerFinder is higher, it is only achieved when it is using evolutionary conservation. However if we do not allow EnhancerFinder to include this data (which is fair, as conservation was part of selection of candidates to the Vista database), its performance is inferior to our method despite much larger number of features.

Feature importance

Great advantage of Random Forest classifiers over SVM methods is the ability to easily analyse their structure and measure the importance of individual features, e.g. single 4-mers. To do that, we used Boruta algorithm ([31]), a wrapper around Random Forest, that runs classification multiple times and compares feature importance defined as loss of accuracy if feature is permuted, with importance of random features (shuffled initial features). It labels features that are significantly more important than the best random feature as ‘Confirmed’, and rejects the features that are less important than best random. The features that do not meet either of the criteria are labeled ‘Tentative’.

Using Boruta algorithm we found features important for 4 mers+Tier1&2 classifier for heart and brain (see summary in Fig. 2b, full data in Additional file 1: Table S2).

Both types of attributes (sequence and histone modification) were found among important ones: most of histone modifications from H1hESC, some histone modifications from HUVEC and other cell types, and almost one-third of 4-mers. While almost all of the ESC histone features were important for both heart and brain, majority of non-ESC modifications were only predictive in specific tissues. The *k*-mers were also much more specific in their relevance.

This is consistent with the view that the basic enhancer markers such as H3K4me1 are actually already deposited on enhancers very early in the stem cell stage and the later modifications of these marks are not adding more information for the classifier. However, the early marks may be helpful in finding the negative examples of sequences with the right features, which are positioned in the chromatin context not allowing them to be activated and therefore not marked epigenetically already in the ESC stage. This is consistent with previous reports we found in *Drosophila* developmental enhancers [18].

Whole-genome predictions correlate with DHS

Activity of an enhancer is dependent on (among other factors) accessibility of DNA in the region where it is located. DNaseI digestion is one of the main methods used to evaluate this property [21]. Using such measurements, allows us to test our prediction quality on a technically independent experimental dataset, as well as to test (at least to some extent) the true negative predictive value (this

is not possible on the randomized training set as we are not sure how many of the random sequences are indeed non-functional. We expect that active enhancers will all be located in open chromatin, although enhancers will consist only part of all DNaseI hypersensitive regions.

We used our classifiers to compute predictions on the whole human genome, divided into overlapping windows of length 1500 bp. We compared values returned by our 4-mers classifiers for windows overlapping DNase Hypersensitive Sites (DHS) and windows without hypersensitivity (non-DHS) (see Fig. 2c and d). For general set of DHS from the ENCODE [34], derived from multiple cell-lines and not specific to heart or brain, the classifiers results for DHS were slightly, but significantly above results for non-DHS (Mann-Whitney tests p -value $< 10^{-4}$). Situation was different with fetal heart and brain DNase data from the Epigenomics Roadmap project – our classifier returned clearly higher rates for DHS in compare to specific-non-DHS (p -values $< 10^{-130}$).

Tissue-specificity of predictions is low

Even though tissue-specific classifiers seem to work well in predicting heart or brain enhancers against random sequences, they perform worse on enhancers from different tissue (that show now activity in selected tissue), although their confidence is usually lower than those from relevant enhancers (see Fig. 3a and b). It is also clearly visible in cross-comparison using DNase-Seq data – predictions for DHS windows specific only to heart or brain (windows with promoters or hypersensitive in both tissues excluded) by tissue-specific classifier is small, but significant. For brain classifier see Fig. 3c (Mann-Whitney test p -value $< 10^{-18}$ for DHS), for heart classifier Additional file 1: Figure S5) (p -value < 0.015).

This result is in agreement with the previous reports by Erwin et al., who also noticed significant overlaps in genome wide predictions made by classifiers trained on data from different tissues. While their approach is to use

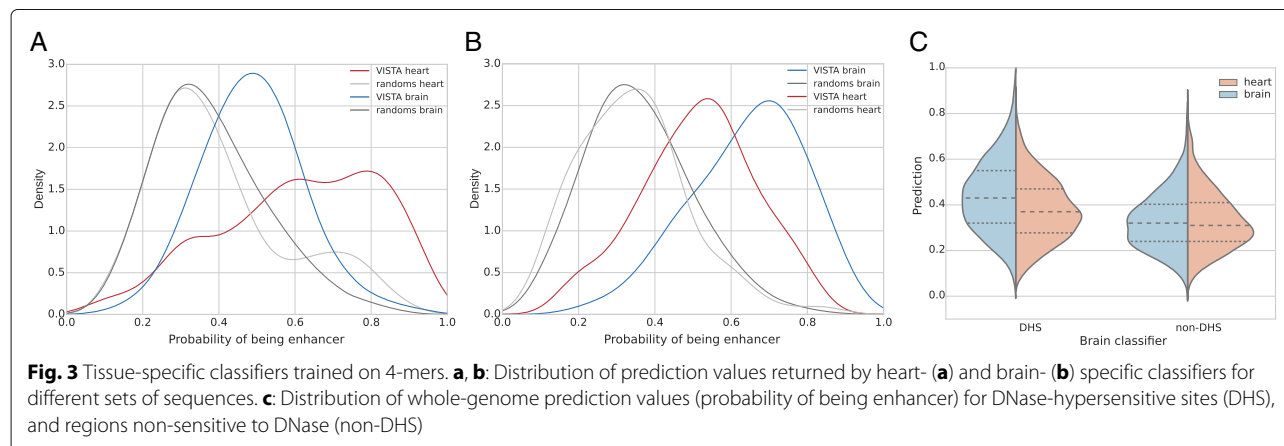
direct machine learning to discern between the known classes of enhancers, we have gone a different route and tried to find a more direct explanation for this problem and a slightly different solution.

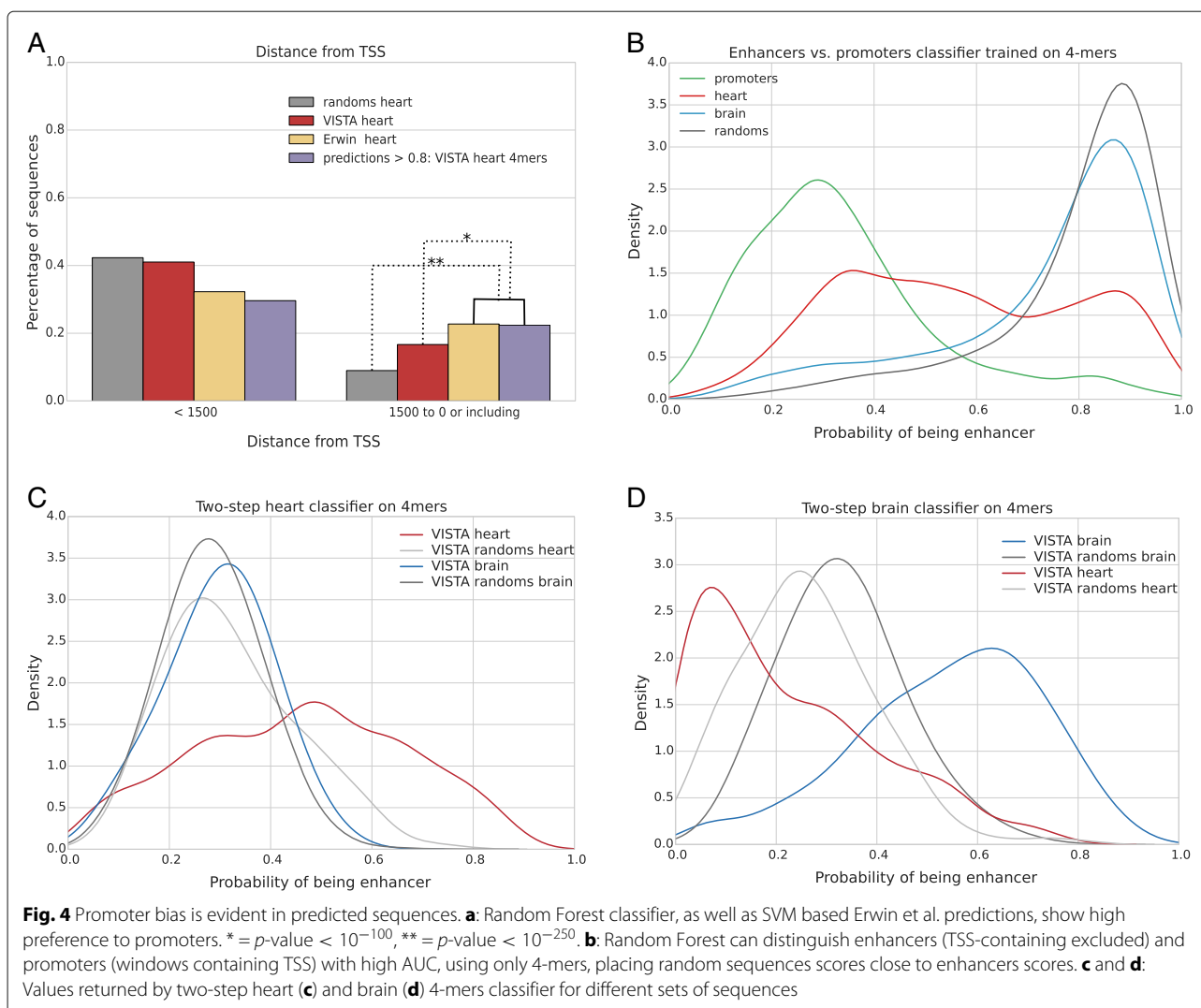
Whole-genome predictions show promoter-bias

We analyzed locations of high confidence enhancer predictions (i.e. windows with score over 0.8) with respect to nearest transcription start site (TSS) (see Fig. 4a). We found that out of all 18,376 windows predicted by 4-mers heart classifier 22% were located within 1500 bp upstream of TSS (TSS-proximal), and for 4807 predicted brain enhancers the ratio was 26%.

This is in contrast with our set of random training sequences where promoter proximal sequences amounted to less than 10% (9% of random sequences for heart classifier and 7% of sequences for brain classifier). This is a significant enrichment (binomial test p -value $< 10^{-250}$). Region of 1500 bases upstream of TSS should contain promoter sequences and many promoter-related transcription factor binding sites, and this promoter-related signal seems to be picked up by the machine learning algorithm leading to a non-specific bias. This problem is not only affecting our method – it is also an issue in EnhancerFinder [22], for which 23% of heart predictions and 16% of brain predictions are TSS-proximal (p -value $< 10^{-100}$). It is related to the fact that the training sets, especially heart-specific sequences are slightly enriched in regions overlapping promoters (17% for heart, 9% for brain, p -values 0.03 and 0.08), however the machine-learning predictions are yet significantly enriched with TSS-proximal regions over the positive examples (p -value $< 10^{-80}$).

Although promoters are well annotated, so TSS-close predictions can easily be filtered out, we were interested whether we could train our model to discern between enhancers and promoters. If successful, this could help us in the task of predicting tissue-specific enhancer





predictors that are not dependent on the knowledge of the “unwanted” enhancer classes, but rather leverage the removal of the promoter bias.

Two-step classifiers improve specificity

Our first approach was to add enhancers vs. promoters classifiers. As a negative training set we used promoter-containing windows selected by our first-level 4-mers classifiers (heart and brain) with prediction over 0.8. As a positive set we used our previous set of heart and brain enhancers, after adjusting lengths of this sequences to our window length, and removing regions containing TSS. While we were able to discern the enhancers from promoters (AUC = 84%), it proved to be unusable for our purposes as still mixed the promoter and enhancer signatures into a single model. That resulted in the situation, where many random sequences were rated high because of their dissimilarity to the set of promoter sequences (See Fig. 4b).

For this reason we have turned to training our second classifier on promoter vs. randomly selected sequences. Here we consider random sequences (of length 1500) the positive training set, and promoter windows the negative set. We combine two classifiers by multiplying their predictions, where a high score can be achieved only if the sequence has similar features to the enhancers from the positive set of the first classifier and dissimilar to the promoters of the second classifier. This multiplication of scores leads to slight decrease of the model performance (e.g. AUC from 0.91 to 0.82 for the brain) likely because some of the enhancers indeed include promoter-like features and are likely to be not specific to only one tissue. Nonetheless, we then show that two-step classifiers for both heart and brain indeed show specificity in their prediction against the other class (See Fig. 4c, d). This is, importantly, despite the fact that the second classifier is not built to specifically exclude the other known tissue, but rather to exclude the non-specific promoter signal.

Discussion and conclusion

Computational predictions of tissue-specific enhancer activity based on the sequence and epigenetic features is an important field of research, given the complexity of metazoan organisms and the difficulty of obtaining comprehensive experimental measurements of such activity.

Given the current wealth of experimental data, this problem can be formulated as a supervised machine learning task with the positive set taken from an experimental dataset such as the Vista database and the negative set usually taken from a controlled randomized set of genomic sequences.

In our paper, we describe a new approach that uses random forests for this classification task. It has several advantages over the previous studies in this area. In particular, we were able to assess the relative utility of different histone modifications as well as different sequence features for prediction of active enhancers in different tissues.

This allowed us to define a greatly reduced set of features including only histone modifications from the embryonic stem cells (8 ChIP-Seq experiments) and k -mers to achieve over 90 percent AUC score.

Using the Boruta package, we were also able to verify which sequence features were the most important and see that the data we have are consistent with the hypothesis that the tissue-specific activity of an enhancer is a combined result of the epigenetic context laid out early in the development and the sequence specific binding of the transcription factors expressed in a given tissue.

We have assessed the genome-wide specificity of such classifiers on the brain and heart related datasets from the Vista database and found that while there is a detectable difference between the positive sequence sets for different tissues, both classifiers are ranking the positive sequences from a different tissue significantly better than the control sequences. This is in agreement with our comparisons with the DNase-seq data, which point to the same conclusion of the predictions being specific only in comparison with negative controls, but not between classifiers trained on different positive sets.

We found that at least part of this problem is related to the fact that such classifiers are prone to “learning” the enrichment of enhancer-proximal sequences in the positive training sets. This leads to great over-representation of promoter sequences in the predictions of both our classifier as well as the previously published methods.

To find out whether the cause of this observation can be the actual set of sequence features contained in promoter-proximal sequences we have tried to construct a classifier based on purely sequence features that would discern promoters from a control set of sequences. This has proven to be possible and indeed we have further shown that combined classifier using both the promoter related features

and the enhancer-trained classifier, is resulting in a much greater specificity between tissues.

We consider our results promising, in the sense that they bring a finer understanding of the mechanisms behind tissue-specific enhancer activity while giving us at the same time useful classifiers and genome-wide predictions. Simultaneously, our results bring new questions into the field. In particular, if our results are correct, we should pay much more attention to the mechanistic difference between enhancers and promoters when building computational methods of analysis and detection of regulatory sequences.

In the long term, these results can lead to better understanding of how mutations in regulatory sequences can disrupt enhancer or promoter function and allow us to better understand the root causes of some genetic diseases.

Additional file

Additional file 1: Supplementary Tables and Figures. This file contains additional tables and figures, such as table of datasets used for training, feature importance table and predictions for recently added VISTA sequences. (PDF 236 kb)

Acknowledgments

Not applicable.

Funding

Publication of this article has been funded by Polish National Science Center grant number **DEC-2014/12/W/NZ1/00463**. This work was partially supported by the Polish National Science Center grants with decision numbers **DEC-2012/05/B/NZ2/00567** and **DEC-2014/12/W/NZ1/00463**.

Availability of data and materials

The software needed to train the models is available at http://github.com/regulomics/enhancer_prediction and the predictions themselves are available at <http://regulomics.mimuw.edu.pl:8888>.

Authors' contributions

BW and JHI have prepared data, trained all classifiers and prepared manuscript. BW and MW have analyzed predictions in context of DHS sites and created website providing predictions. All authors read and agreed upon the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

About this supplement

This article has been published as part of *BMC Medical Genomics* Volume 10 Supplement 1, 2017: Selected articles from the 6th Translational Bioinformatics Conference (TBC 2016): medical genomics. The full contents of the supplement are available online at <https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-10-supplement-1>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published: 24 May 2017

References

- Marsman J, Horsfield JA. Long distance relationships: enhancer–promoter communication and dynamic gene transcription. *Biochim Biophys Acta (BBA) - Gene Regul Mech.* 2012;1819(11–12): 1217–27. doi:10.1016/j.bbagrm.2012.10.008.
- Wilczynski B, Furlong EEM. Challenges for modeling global gene regulatory networks during development: insights from *Drosophila*. *Dev Biol.* 2010;340(2):161–9. doi:10.1016/j.ydbio.2009.10.032. Accessed 29 Mar 2016.
- Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, Bergman CM, Halfon MS. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. *Nucleic Acids Res.* 2011;39(Database issue):118–23. doi:10.1093/nar/gkq999. Accessed 5 Jan 2011.
- Kvon EZ, Kazmar T, Stampfel G, Yanez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, Stark A. Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nat Adv Online Publ.* 2014. doi:10.1038/nature13395. Accessed 27 June 2014.
- Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518(7539):317–30.
- Krivan W, Wasserman WW. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* 2001;11(9):1559. doi:10.1101/gr.180601. Accessed 13 May 2009.
- Wasserman WW, Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol.* 1998;278(1):167–81. doi:10.1006/jmbi.1998.1700. Accessed 13 May 2009.
- Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc Natl Acad Sci USA.* 2002;99(2): 757. doi:10.1073/pnas.231608898. Accessed 12 Mar 2010.
- Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell.* 2006;124(1):47–59.
- Wilczynski B, Dojer N, Patelak M, Tiuryn J. Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs. *BMC Bioinforma.* 2009;10(1):82.
- Arunachalam M, Jayasurya K, Tomancak P, Ohler U. An alignment-free method to identify candidate orthologous enhancers in multiple *drosophila* genomes. *Bioinformatics.* 2010;26(17):2109–15.
- Kleftogiannis D, Kalnis P, Bajic VB. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform.* 2015;101:. doi:10.1093/bib/bbv101. Accessed 19 Jan 2016.
- Wilczynski B, Tiuryn J. Fastbill: An improved tool for prediction of cis-regulatory modules. *J Comput Biol.* 2017;24(3):193–9. doi:10.1089/cmb.2016.0108. <https://www.ncbi.nlm.nih.gov/pubmed/27710048>.
- Dabrowski M, Dojer N, Krystkowiak I, Kaminska B, Wilczynski B. Optimally choosing pwm motif databases and sequence scanning approaches based on chip-seq data. *BMC Bioinforma.* 2015;16(1):1.
- Kazemian M, Zhu Q, Halfon MS, Sinha S. Improved accuracy of supervised crm discovery with interpolated markov models and cross-species comparison. *Nucleic Acids Res.* 2011;39(22):9463–72. doi:10.1093/nar/gkr621. <https://www.ncbi.nlm.nih.gov/pubmed/21821659>.
- Szalkowski AM, Schmid CD. Rapid innovation in chip-seq peak-calling algorithms is outdistancing benchmarking efforts. *Brief Bioinform.* 2011;12(6):626–33.
- Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet.* 2007;39(3):311–8.
- Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczynski B, Riddell A, Furlong EEM. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat Genet.* 2012;44(2):. doi:10.1038/ng.1064. Accessed 10 Jan 2012.
- Wilczynski B, Furlong EEM. Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol Syst Biol.* 2010;6:. doi:10.1038/msb.2010.35. Accessed 22 July 2010.
- Thomas S, Li XY, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, Biggin MD, Stamatoiyannopoulos JA. Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. *Genome Biol.* 2011;12(5):43. doi:10.1186/gb-2011-12-5-r43. Accessed 2011-08-12.
- Mercer TR, Edwards SL, Clark MB, Neph SJ, Wang H, Stergachis AB, John S, Sandstrom R, Li G, Sandhu KS, Ruan Y, Nielsen LK, Mattick JS, Stamatoiyannopoulos JA. DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nat Genet.* 2013;45:. doi:10.1038/ng.2677. Accessed 26 June 2013.
- Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA. Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput Biol.* 2014;10(6): 1003677. doi:10.1371/journal.pcbi.1003677.
- Fletez-Brant C, Lee D, McCallion AS, Beer MA. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res.* 2013;41(W1):544–56. doi:10.1093/nar/gkt519. Accessed 16 Sept 2013.
- Podsiadło A, Wrzesień M, Paja W, Rudnicki W, Wilczyński B. Active enhancer positions can be accurately predicted from chromatin marks and collective sequence motif data. *BMC Syst Biol.* 2013;7(Suppl 6):16.
- Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoiyannopoulos J, Ernst J, Kellis M, Ren B. Rfecs: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol.* 2013;9(3):1002968.
- Firpi HA, Ucar D, Tan K. Discover regulatory dna elements using chromatin signatures and artificial neural network. *Bioinformatics.* 2010;26(13):1579–86.
- Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, Cheung VG, Kraus WL, Lis JT, Siepel A. Identification of active transcriptional regulatory elements from gro-seq data. *Nat Methods.* 2015;12(5):433–8.
- Zhu Y, Sun L, Chen Z, Whitaker JW, Wang T, Wang W. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res.* 2013;41(22):10032–43.
- Bednarz P, Wilczyński B. Supervised learning method for predicting chromatin boundary associated insulator elements. *J Bioinforma Comput Biol.* 2014;12(06):1442006.
- Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983;148(3):839–43.
- Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw.* 2010;36(11):1–13.
- Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32. doi:10.1023/A:1010933404324.
- Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* 2007;35(Database issue):88–92.
- Bernstein BE, Birney E, et al D. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57–74.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit





A New Evolutionary Rough Fuzzy Integrated Machine Learning Technique for microRNA selection using Next-Generation Sequencing data of Breast Cancer

Jnanendra Prasad Sarkar^{*†}

Larsen & Toubro Infotech Ltd.
Pune, India

Indrajit Saha^{*}

Dept. of Computer Science and Engg.,
National Institute of Technical
Teachers' Training and Research
Kolkata, India

Somnath Rakshit^{*}

Centre of New Technologies,
University of Warsaw
Warsaw, Poland

Monalisa Pal

Machine Intelligence Unit,
Indian Statistical Institute
Kolkata, India

Michał Własnowolski[‡]

Faculty of Mathematics and
Information Science, Warsaw
University of Technology
Warsaw, Poland

Anasua Sarkar

Department of Computer Science
and Engineering, Jadavpur University
Kolkata, India

Ujjwal Maulik

Department of Computer Science
and Engineering, Jadavpur University
Kolkata, India

Dariusz Plewczynski[§]

Center of New Technologies,
University of Warsaw
Warsaw, Poland

ABSTRACT

MicroRNAs (miRNA) play an important role in various biological process by regulating gene expression. Their abnormal expression may lead to cancer. Therefore, analysis of such data may discover potential biological insight for cancer diagnosis. In this regard, recently many feature selection methods have been developed to identify such miRNAs. These methods have their own merits and demerits as the task is very challenging in nature. Thus, in this article, we propose a novel wrapper based feature selection technique with the integration of Rough and Fuzzy sets, Random Forest and Particle Swarm Optimization, to identify putative miRNAs that can solve the underlying biological problem effectively, i.e. to separate tumour and control samples. Here, Rough and Fuzzy sets help to address the vagueness and overlapping characteristics of the dataset while performing clustering. On the other hand, Random Forest is applied to perform the classification task on the clustering results

to yield better solutions. The integrated clustering and classification tasks are considered as an underlying optimization problem for Particle Swarm Optimization method where particles encode features, in this case, miRNAs. The performance of the proposed wrapper based method has been demonstrated quantitatively and visually on next-generation sequencing data of breast cancer from The Cancer Genome Atlas (TCGA). Finally, the selected miRNAs are validated through biological significance tests. The code and dataset used in this paper are available online¹.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; • **Applied computing** → **Bioinformatics**; **Transcriptomics**; **Systems biology**;

KEYWORDS

Breast Cancer, Clustering, Fuzzy Set, Feature Selection, Particle Swarm Optimization, Random Forest, Rough Set

ACM Reference Format:

Jnanendra Prasad Sarkar, Indrajit Saha, Somnath Rakshit, Monalisa Pal, Michał Własnowolski, Anasua Sarkar, Ujjwal Maulik, and Dariusz Plewczynski. 2019. A New Evolutionary Rough Fuzzy Integrated Machine Learning Technique for microRNA selection using Next-Generation Sequencing data of Breast Cancer. In *Genetic and Evolutionary Computation Conference Companion (GECCO '19 Companion)*, July 13–17, 2019, Prague, Czech Republic. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3319619.3326836>

1 INTRODUCTION

MicroRNAs (miRNA) are small non-coding molecules of single-stranded RNA, 22-25 nucleotide long. MicroRNAs (miRNAs) bind

¹<http://www.nittrkol.ac.in/indrajit/projects/mirna-pso-rfcm-rf-berastcancer/>

^{*}Equally contributed

[†]Additional Affiliation: Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

[‡]Additional Affiliation: Center of New Technologies, University of Warsaw, Warsaw, Poland

[§]Additional Affiliation: Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6748-6/19/07...\$15.00

<https://doi.org/10.1145/3319619.3326836>

partially with complementary sites in target messenger RNAs (mRNAs) and thus regulate gene expression of animals and plants, where dysregulation causes tumor formations [8]. MicroRNAs are also found to have important roles in other diseases like diabetes [2], infectious disease [22], various neurodegenerative disorder [11]. In recent decade, various analytical processes have been studied on miRNA. Those are identifying set of miRNAs derived from common primary transcripts [21], co-expression analysis between neighbouring miRNAs [4], prediction of miRNA targets [10] and specificity of miRNA in particular tissue [23]. Individual miRNA can target many mRNAs based on sequence complementarity. However, a significant fraction of these interactions may depend on cell type, context [3] and also on the binding of additional co-factors [12]. Moreover, a smaller subset of target interactions actually cause tumour development. Therefore, it is important to identify the potential set of miRNAs, which is very challenging task. Therefore, analysis of miRNAs in various aspects has become major focused area of research in recent decades. Recent studies reveal that some miRNAs are differently expressed in both normal and cancerous tumour tissues of all types. Additionally, it is also seen that some miRNAs are differently expressed in specific tumour tissue. So, it suggests that there might be any link between miRNAs and oncogenesis. Also, diagnosis of cancer might be possible from onco-miRNA signature. Therefore, in addition to wet laboratory experiments, computational methods can also be useful to detect onco-miRNA signature and an alternative method for medical diagnosis. In this regards, different machine learning techniques like K -Nearest Neighbour (K -NN) [1], Support Vector Machine (SVM) [7], Decision tree (DT) [20], Naive Bayesian classifier (NB) [9] etc. are used for analysis. However, performance of these heterogeneous methods depends very much on the selection of features. This fact motivated us to propose a novel method for identifying potential set of features.

In this regard, it is important to address inherent vagueness, uncertainty and overlapping characteristics within the dataset. Fuzzy C-Means (FCM) [5] using Fuzzy set theory can handle overlapping characteristics. However, it is very sensitive to noisy data. Thus, variants of FCM [16, 19] have been developed for the same to handle subtle vagueness and uncertainty by incorporating Rough set theory [18] and known as Rough Fuzzy C-Means (RFCM) [16]. According to Rough set theory, a point can either belong to a particular cluster with membership value 1 or to the boundary region of multiple clusters. The boundary region is considered as overlapping region of more than one clusters. Hence, we have used both Rough and Fuzzy set theories together to handle vagueness, uncertainty and overlapping characteristics of the dataset. However, Rough Fuzzy integrated technique yields clusters having set of crisp and rough points. Therefore, to get better clustering results, well-known machine learning method called Random Forest (RF) [6] is applied on rough points after being trained on crisp points. The integrated clustering (RFCM) and classification (RF) tasks are considered as an underlying optimization problem for Particle Swarm Optimization (PSO) [14] in order to identify potential set of features, in this case miRNAs, to perform better separation of tumour and control samples. Here PSO encodes miRNAs as elements of a particle. The proposed wrapper based technique is abbreviated as PSO-RFCM-RF.

The publicly available Breast Invasive Carcinoma (BRCA) dataset² is used to demonstrate the performance of proposed method. The breast cancer is believed to be most widely diagnosed cancer type and mostly found within female population. In female body, it mostly begins in cells of the lobules which are known as milk-producing glands. Thereafter, it might get spread outside milk ducts. Unlike non-invasive, invasive cancers grow into healthy tissues. Sometimes, both non-invasive and invasive cancers are found in same specimen. Even today, the exact cause for this cancer is not fully known, whereas, the proper analysis of various biomolecules may bring more insights. Thus, the analysis of miRNA expression and its proper selection may help to achieve that goal. In this regard, the proposed PSO-RFCM-RF is used and the selected miRNAs are validated quantitatively as well as through biological significance tests.

2 EVOLUTIONARY ROUGH FUZZY INTEGRATED MACHINE LEARNING TECHNIQUE

This section describes the proposed wrapper based feature selection technique.

Algorithm 1 Steps of the RFCM

Input:
 X , the dataset
 η , the fuzzy exponent
 ϵ , a small real threshold value between [0,1]
 K , the number of cluster
 f_{LW} , relative weight for lower approximation of rough clustering, $0 < f_{LW} < 1$
Output: $[\mu]$ where, $1 \leq l \leq K$ and $1 \leq i \leq n$

```

1: Select random  $K$  points from dataset as  $K$  cluster means
2: repeat
3:   Compute  $\mu_{li}$  for all  $n$  points using Equation 3
4:   Compute the difference between highest two computed membership,  $\mu_{li}$  of each and every  $n$  data points
      // Let  $\mu_{li}$  and  $\mu_{hi}$ , highest and second highest computed membership values of  $x_i$  among all  $K$  clusters, where  $1 \leq l, h \leq K$  and  $h \neq l$ 
5:   Compute the value of threshold  $\Delta$ 
      //  $\Delta$  is the mean of  $(\mu_{li} - \mu_{hi})$ ,  $\forall i = 1, 2, \dots, n$ 
6:   if  $(\mu_{li} - \mu_{hi}) > \Delta$  then
7:      $\mu_{li} \leftarrow 1$ ,  $\mu_{hi} \leftarrow 0 \forall h = 1, 2, \dots, K$  and  $h \neq l$ 
      //  $x_i$  is exactly classified to  $\underline{B}(C_l)$ , also to  $\bar{B}(C_l)$  as per RST
8:   else
9:     Keep  $\mu_{hi}$  unchanged  $\forall h = 1, 2, \dots, K$ 
      //  $x_i$  can belong to Upper Approximation of multiple clusters. Hence,  $x_i$  belong to  $\bar{B}(C_l)$  and  $\bar{B}(C_h)$ 
10:  end if
11:  Compute new mean with the help of Equation 4
12: until  $|Current J_{RFCM} - Previous J_{RFCM}| \leq \epsilon$ 
13: return  $[\mu]$  where,  $1 \leq l \leq K$  and  $1 \leq i \leq n$ 

```

The proposed clustering and classification integrated wrapper based feature selection technique uses Rough and Fuzzy sets to cluster a dataset $X = \{x_i \mid 1 \leq i \leq n\}$. The steps of clustering are described in Algorithm 1, where it produces crisp and rough sets of points. Crisp set of points are crisply classified into *lower approximation* region whereas rough points belong to boundary region of multiple clusters. According to Rough set theory [18], *lower approximation* ($\underline{B}(X)$) and *upper approximation* ($\bar{B}(X)$) are

²<https://cancergenome.nih.gov/>

Algorithm 2 Steps of RFCM-RF

Input:
 X , the dataset
 η , the fuzzy exponent
 ϵ , a small real threshold value between [0,1] used to terminate RFCM
 K , the number of cluster
 f_{LW} , relative weight for lower approximation of rough clustering, $0 < f_{LW} < 1$
 T , Number of tree for RF
Output: F , the final class label vector of X

- 1: Using Algorithm 1 produce crisp dataset, $\mathbb{L} = \{x_i \in \underline{B}(C_l) \mid 1 \leq l \leq K \text{ and } 1 \leq i \leq n\}$ and corresponding cluster label vector, λ_1
- 2: Classify $\mathbb{L}^* = (X - \mathbb{L})$ using RF, trained by \mathbb{L} and λ_1 to get label vector, λ_2
- 3: Combine λ_1 and λ_2 to get final cluster label vector, F , where F should be in order of X
- 4: **return** F

defined in Equation 1, where U is non-empty set called *universe* and B determines the *equivalence* or *indiscernibility* relation. An *indiscernibility* class containing x is denoted as $B(x)$. The difference between *upper* and *lower approximation* regions, (i.e., $BN(X) = \overline{B}(X) - \underline{B}(X)$), is called boundary region of X . If $BN(X)$ is empty then X is called crisp set of points, otherwise it is called as rough set of points.

$$\underline{B}(X) = \bigcup_{x \in U} \{B(x) \mid B(x) \subseteq X\}; \quad \overline{B}(X) = \bigcup_{x \in U} \{B(x) \mid B(x) \cap X \neq \emptyset\} \quad (1)$$

Algorithm 1 optimises the objective function as defined in Equation 2.

$$J_{RFCM} = \begin{cases} f_{LW} \times \mathcal{A} + f_{BN} \times \mathcal{B}, & \text{if } \underline{B}(C_l) \neq \emptyset, BN(C_l) \neq \emptyset \\ \mathcal{A}, & \text{if } \underline{B}(C_l) \neq \emptyset, BN(C_l) = \emptyset \\ \mathcal{B}, & \text{if } \underline{B}(C_l) = \emptyset, BN(C_l) \neq \emptyset \end{cases} \quad (2)$$

$$\mathcal{A} = \sum_{l=1}^K \sum_{x_i \in \underline{B}(C_l)} (\mu_{li})^\eta D(c_l, x_i); \quad \mathcal{B} = \sum_{l=1}^K \sum_{x_i \in BN(C_l)} (\mu_{li})^\eta D(c_l, x_i)$$

$$\mu_{li} = \frac{1}{\sum_{h=1}^K \left(\frac{D(c_l, x_i)}{D(c_h, x_i)} \right)^{\frac{2}{\eta-1}}}; \quad \sum_{l=1}^K \mu_{li} = 1 \text{ for } 1 \leq l \leq K; \quad 1 \leq i \leq n, \quad (3)$$

where, C_l is l th cluster and $D(c_l, x_i)$ measures Euclidean distance of the point, x_i from the center of cluster, $c_l \in C_l$. η is weighting coefficient while μ_{li} as defined in Equation 3 represents the fuzzy membership value or the degree of belongingness of the i th point to the l th cluster. According to rough set theory, the degree of belongingness of the points is 1 for a particular cluster within *lower approximation* region. Therefore, Equation 2 can be rewritten as $\mathcal{A} = \sum_{l=1}^K \sum_{x_i \in \underline{B}(C_l)} D(c_l, x_i)$. The cluster center is updated by Equation 4.

$$c_l = \begin{cases} f_{LW} \times \mathcal{G}_{LW} + f_{BN} \times \mathcal{G}_{BN}, & \text{if } \underline{B}(C_l) \neq \emptyset, BN(C_l) \neq \emptyset \\ \mathcal{G}_{LW}, & \text{if } \underline{B}(C_l) \neq \emptyset, BN(C_l) = \emptyset \\ \mathcal{G}_{BN}, & \text{if } \underline{B}(C_l) = \emptyset, BN(C_l) \neq \emptyset \end{cases} \quad (4)$$

where,

$$\mathcal{G}_{LW} = \frac{\sum_{x_i \in \underline{B}(C_l)} x_i}{|\underline{B}(C_l)|}; \quad \mathcal{G}_{BN} = \frac{\sum_{x_i \in BN(C_l)} \{(\mu_{li})^\eta\} x_i}{\sum_{x_i \in BN(C_l)} \{(\mu_{li})^\eta\}}$$

However, using RFCM, it is difficult to determine the definite belongingness of rough points in a particular cluster. Thus, Random Forest (RF) is used to classify those rough points with the help of crisp points that are used to train the RF. It refines the performance of the clustering. The steps of RFCM-RF are described in Algorithm 2. Furthermore, the RFCM-RF is considered as an optimization problem for Particle Swarm Optimization (PSO) while identifying the potential set of features, in this case miRNAs. Here PSO is used as a global optimizer to achieve the optimal solution, in this case, the set of miRNAs while performing clustering and classification tasks in integrated fashion, i.e. RFCM-RF. The method is described in Algorithm 3.

Algorithm 3 Steps of PSO integrated RFCM-RF

Input:
 X , the dataset
 η , the fuzzy exponents
 ϵ , a small real threshold value between [0,1] used to terminate RFCM
 K , the number of cluster
 f_{LW} , relative weight for lower approximation of rough clustering, $0 < f_{LW} < 1$
 T , Number of tree for RF
 N_{par} , Number of particles
 N_{itr} , Number of iteration for PSO
 L , Length of particle
 α , Inertia weight $\in [0.5, 1]$
 β_1, β_2 , Cognitive and Social constant
Output: S_{best} , Best feature subset

- 1: $\hat{X} \leftarrow Preprocess(X)$
- 2: $\mathcal{P}^{(t)} \leftarrow InitialPopulation(\hat{X}, N_{par}, L)$
- 3: **for** $i = 1$ to N_{itr} **do**
- 4: $[\mathcal{P}_{l_{best}}^{(t)}, \mathcal{P}_{g_{best}}^{(t)}] \leftarrow FitnessRFCMRF(\hat{X}, \mathcal{P}^{(t)})$
- 5: $\mathcal{V}^{(t+1)} \leftarrow Velocity(\mathcal{P}^{(t)}, \mathcal{V}^{(t)}, \mathcal{P}_{l_{best}}^{(t)}, \mathcal{P}_{g_{best}}^{(t)})$ // using Equation 5
- 6: $\mathcal{P}^{(t+1)} \leftarrow Position(\mathcal{V}^{(t+1)}, \mathcal{P}^{(t)})$ // using Equation 6
- 7: $S_{best} \leftarrow BestFeatureSet(\mathcal{P}^{(t)}, \mathcal{P}_{g_{best}}^{(t)})$
- 8: **end for**
- 9: **return** S_{best}

PSO works with a population of candidate solution called Swarm where candidate solutions are represented as particles (\mathcal{P}_j , where $j = 1, 2, \dots, N_{par}$ and N_{par} is number of particles). Element of each such particle is composed of position and length (\mathcal{L}). The movement of a particle is tracked by updating velocity (\mathcal{V}_j) and position as defined in Equation 5.

$$\mathcal{V}_j^{(t+1)} = \alpha \times \mathcal{V}_j^{(t)} + \beta_1 \times (\mathcal{P}_{l_{best}}^{(t)} - \mathcal{P}_j^{(t)}) + \beta_2 \times (\mathcal{P}_{g_{best}}^{(t)} - \mathcal{P}_j^{(t)}) \quad (5)$$

$$\mathcal{P}_j^{(t+1)} = \mathcal{P}_j^{(t)} + \mathcal{V}_j^{(t+1)} \quad (6)$$

Where, t is time of different iterations, α is the inertia weight $\in [0.5, 1]$, β_1 is cognitive constant and β_2 is social constant. Moreover, $\mathcal{P}_{l_{best}}$ and $\mathcal{P}_{g_{best}}$ represent local best particle of current iteration and global best particle till current iteration respectively. PSO algorithm terminates after fix number of iterations. In *InitialPopulation* step, a particle is prepared after random selection of elements (in this case miRNAs) from pre-processed dataset. The encoded particle is then used to compute fitness using objective function mentioned in Algorithms 1 and 2. The fitness value ranges from 0 to 100 where, higher value denotes better result. Based on fitness value, local and global best particles are identified to update the *Velocity*. Thereafter, new position of the particle is computed using updated velocity. Finally, the algorithm gets terminated after a fix number of iterations

producing the optimal feature set. The entire process is run for 50 times in our experiment.

The proposed technique, PSO-RFCM-RF is random in nature, which has a probability of having false positive or false negative while selecting miRNAs. To reduce this probability of false positive or false negative, PSO-RFCM-RF is executed for 50 times followed by ranking of miRNAs based on occurrence in 50 different sets of features. Maximum number of occurrence of any miRNAs over 50 runs indicates that it is significant for producing the better fitness value by reducing error while assigning points to a cluster. After 50 runs, PSO-RFCM-RF ensures that miRNAs are ranked according to their occurrence and finally set of miRNAs are selected from sorted list. Here the number of runs is an important factor for reducing false negative. Mathematically, it has been found that 50 runs are sufficient to reduce false negative. Suppose, at each run, PSO-RFCM-RF selects random $\mathbb{S} = 10$ miRNAs from the entire collection of miRNAs, \mathbb{D} and the total number of runs = \mathbb{R} . For $i = 1, 2, \dots, \mathbb{D}$, let \mathbb{V}_i is the Bernoulli distributed indicator variable where $\mathbb{V}_i = 1$ if miRNA, m_i never gets selected. The probability of selecting m_i in a single run is $= \mathbb{S}/\mathbb{D}$ and probability that it does not get selected is $= (1 - \mathbb{S}/\mathbb{D})$. Hence the expectation of \mathbb{V}_i can mathematically be defined as in Equation 7.

$$\mathbb{E}[\mathbb{V}_i] = Pr(\mathbb{V}_i) = (1 - \frac{\mathbb{S}}{\mathbb{D}})^{\mathbb{R}} \quad (7)$$

Let us assume, $\mathbb{V} = \sum_{i=1}^{\mathbb{D}} \mathbb{V}_i$ is the random variable which counts the number of miRNAs that do not belong to the final set of the miRNAs at least once. By linearity relation of the expectation, Equation 8 can be written as below.

$$\mathbb{E}[\mathbb{V}] = \sum_{i=1}^{\mathbb{D}} \mathbb{E}[\mathbb{V}_i] = \mathbb{D} * (1 - \frac{\mathbb{S}}{\mathbb{D}})^{\mathbb{R}} \quad (8)$$

Therefore, it can be written that,

$$\mathbb{E}[(\mathbb{D} - \mathbb{V})] = \mathbb{D} - \mathbb{E}[\mathbb{V}] \quad (9)$$

Substituting the parameters \mathbb{S} , \mathbb{R} and \mathbb{D} with the values as 10, 50 and 244 respectively, the expected number of miRNAs reported at least once after 50 runs is 212 and the expected number of new miRNAs added in a further iteration is 1. However, in our experiment, the sorted number of miRNAs is 60. Hence, it is proved that 50 runs are sufficient to get a stable set of miRNAs to reduce the probability of false negative. This justifies the process of selection of miRNAs and determining 50 runs in the proposed technique.

3 COMPLEXITY ANALYSIS

3.1 Space Complexity Analysis

PSO-RFCM-RF mostly needs space to store data, population, centers of the K clusters, fuzzy membership matrix and fitness value of each particle of population. Additional space is required for processing of RF. Therefore, overall space complexity can be computed as $O(nm + Km + Kn + \mathbb{T}nm + KmN_{par})$. After simplifying, the worst case space complexity is $O(n^2)$ when $n = m$.

3.2 Time Complexity Analysis

Worst case time complexity of PSO-RFCM-RF mainly depends on two parts, (a) time to compute objective function, which is time

complexity of RFCM-RF and (b) time required for standard PSO algorithm. For first part, majority of the time, RFCM-RF spends on computation of fuzzy membership matrix, searching for two highest two membership values, computation of centers of each cluster and additionally for RF processing. Considering all these activities, the overall time complexity can be derived as $O(4Knm + Kn + \mathbb{T}nm\log(n))$. For second part, PSO takes time overall as $O(n\log(n)N_{par})$ for each iteration. Therefore, time complexity of PSO-RFCM-RF can be considered as $O(4Knm + Kn + \mathbb{T}nm\log(n)) + n\log(n)N_{par}$. After simplification, worst case time complexity is $O(n^2)$, when $n = m$, for single iteration.

Table 1: Statistics of Patients in BRCA data

Data Category	Number of Patients	Avg. Age of Patient (in years)	Avg. days to followup
Tumour	762	57.98	1288.29
Control	87	58.64	835.17

Table 2: Top 10 miRNAs with their up/down regulation, p-value and PubMed ID

miRNA	Regulation (Up/Down)	p-value	PubMed ID
hsa-mir-139	Down	1.92e-50	26497851
hsa-mir-21	Up	3.25e-49	29552160
hsa-mir-183	Up	4.17e-47	26170234
hsa-mir-96	Up	7.57e-47	24366472
hsa-mir-486	Down	3.36e-41	25027758
hsa-mir-10b	Down	1.43e-47	16103053
hsa-mir-145	Down	1.49e-46	25124875
hsa-mir-144	Down	4.87e-32	29387244
hsa-mir-15a	Up	5.23e-20	28979704
hsa-mir-182	Up	8.31e-44	19574223

Table 3: Results produced by different feature selection methods using 10-folds cross-validation on BRCA data

Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F-measure (%)
PSO-RFCM-RF	93.25	89.35	93.02	93.35	91.15
SNR-RF	76.39	84.21	78.98	74.00	74.39
t-test-RF	76.82	84.64	79.41	74.43	74.82
RankSum-RF	76.39	84.21	78.98	74.00	74.39
JMI-RF	75.74	85.27	77.59	72.25	74.32
mRMR-RF	75.74	85.27	77.59	72.25	74.32
MIFS-RF	76.67	85.61	79.00	74.33	74.69

4 EXPERIMENTAL RESULTS

4.1 Dataset Preparation

We have used NGS based miRNA expression data of Breast Invasive Carcinoma (BRCA) from The Cancer Genome Atlas (TCGA)³

³<https://cancergenome.nih.gov/>

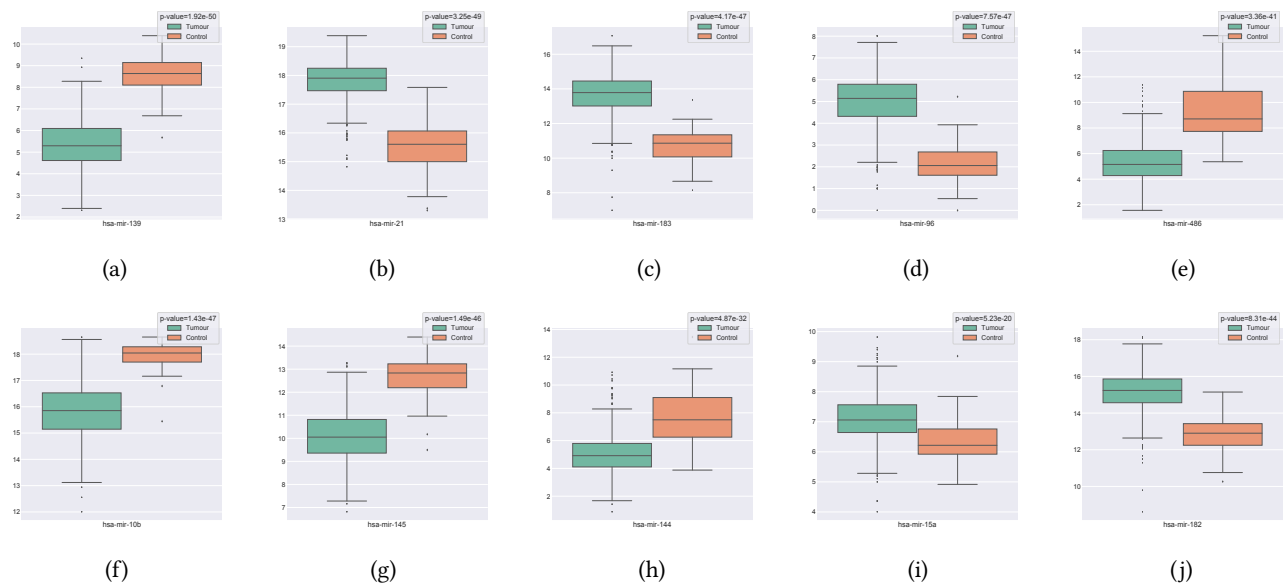


Figure 1: Box plots showing the change in expression values for the selected top 10 miRNAs identified by PSO-RFCM-RF, (a) hsa-mir-139 (b) hsa-mir-21 (c) hsa-mir-183 (d) hsa-mir-96 (e) hsa-mir-486 (f) hsa-mir-10b (g) hsa-mir-145 (h) hsa-mir-144 (i) hsa-mir-15a (j) hsa-mir-182

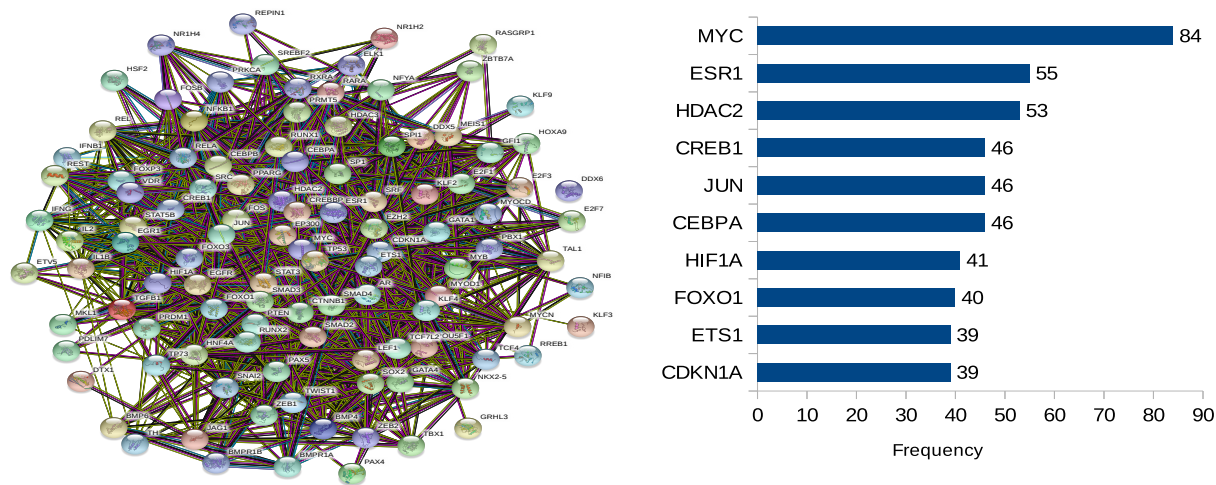


Figure 2: PPI Network for proteins (TFs) that target the top 10 miRNAs, as obtained from the TransmiR Database and the barplot shows degree of top 10 connected proteins

prepared by Illumina sequencing technology where the expression value has been computed in form of reads per million count (RPM). The dataset contains 1046 miRNA expression values for 762 tumour patients and 87 control data as shown in Table 1. Control data comprises patients who are not affected by cancer. Each patient is encoded with barcode like “TCGA-S3-A6ZH-01A-22R-A32K-13”. Barcode is read as “TCGA”: Project, “S3”: Tissue source site (TSS),

“A6ZH”: Participant, “01”: Sample type; “A”: Vial, “22”: Order of portion; “R”: Molecular type of analyte, “A32K”: Plate, “13”: Center. A few preprocessing activities have been performed for the dataset before using in experiments. In the collected dataset, there are many miRNAs which have zero expression values. Thus, such miRNAs are excluded from the dataset which in turn, reduces the number of miRNAs from 1046 to 244. Moreover, the expression values of miRNAs are also normalized by log function with base 2.

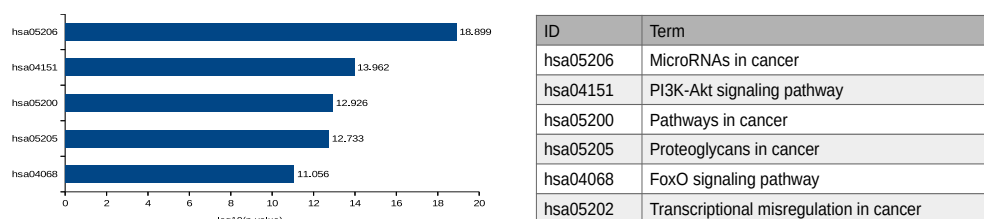


Figure 3: Bar plot of the significant KEGG Pathways for selected top 10 miRNA

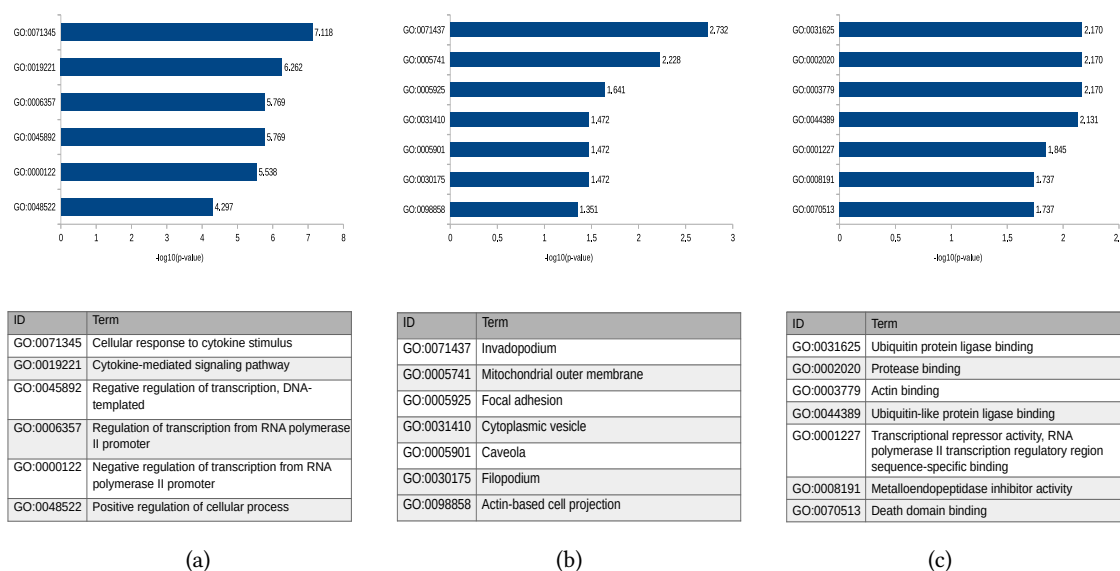


Figure 4: Bar plot of GO Enrichment analysis for, (a) Biological Process (b) Cellular Component and (c) Molecular Function of selected top 10 miRNA identified by PSO-RFCM-RF

4.2 Input Parameters and Performance Metrics

Input parameter values are set experimentally and those are Fuzzy Exponent, $\eta = 2$; Number of particles, $N_{par} = 50$; Number of Iterations, $N_{itr} = 50$; Length of particles, $L = 10$; Cognitive constant, $\beta_1 = 2$; Social constant, $\beta_2 = 2$; Inertia Weight, $\alpha = 0.9$; Number of trees for RF, $T = 50$; Relative weight for lower approximation of RST, $\omega_{low} = 0.95$ and boundary approximation, $\omega_{up} = 0.05$. All the algorithms have been implemented in Matlab and executed on an Intel Core i5-2410M CPU at 2.30 GHz Machine with 8GB RAM and Windows 7 operating system. Moreover, the proposed technique is validated using statistical metrics like *Accuracy*, *Precision*, *Sensitivity*, *Specificity* and *F-measure* respectively.

4.3 Results

PSO-RFCM-RF technique executes 50 times, where PSO maintains the population size as 50. Each particle in the population of PSO denotes a possible solution. To evaluate each such particle, RFCM-RF method is applied to compute fitness value. In each particle, randomly 10 elements, in this case miRNAs, are selected for evaluating

the fitness. Based on the fitness value, local and global best solutions are identified. Such 50 global best solutions are considered after 50 individual run, where each solution contains 10 miRNAs. Based on the occurrence of each miRNA in 50 solutions, top 10 miRNAs are selected and reported in Table 2. Thereafter, these miRNAs are used to perform the classification task using RF with 10-fold cross validation and the results are reported in Table 3. Here, the selection of top 10 miRNAs has been done in order to avoid the false negative. Moreover, it is found from Table 3, PSO-RFCM-RF produces average percentage values of Accuracy, Precision, Sensitivity, Specificity and F-measure as 93.25, 89.35, 93.02, 93.35 and 91.15 respectively on such 10 miRNAs better as compared to the other well-known feature selection techniques viz. Signal-to-Noise Ratio (SNR), t-test, RankSum, Joint Mutual Information (JMI), Minimum Redundancy Maximum Relevance (mRMR) and Mutual Information-based Feature Selection (MIFS) which have been applied on top 10 miRNAs as identified by them. Apart from this, the Figure 1 shows the change of expression of selected top 10 miRNAs using box plot and the corresponding *p-value* is reported after performing Kruskal-Wallis

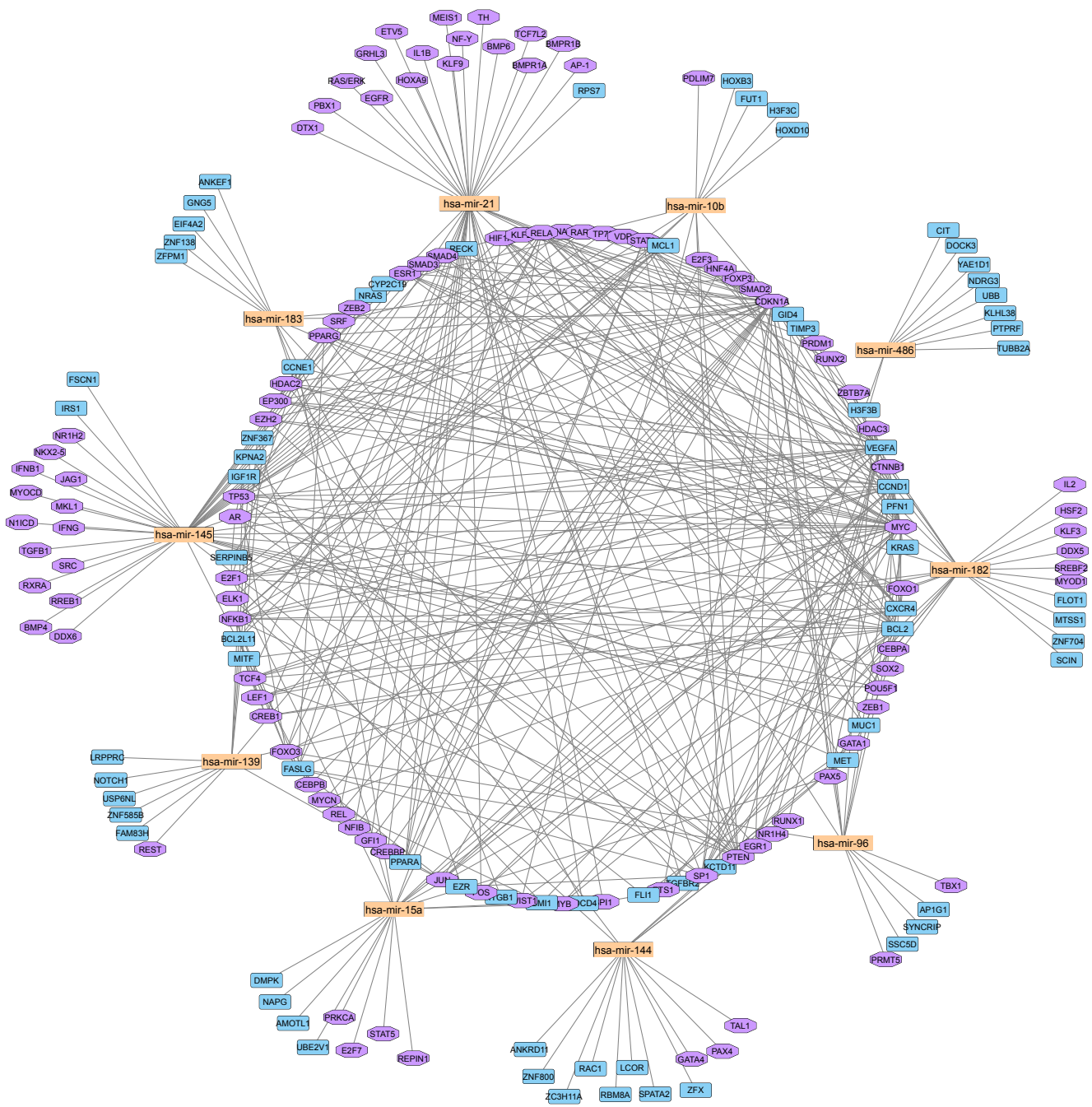


Figure 5: Network Plot of selected top 10 miRNAs that are associated with genes and TFs (Nodes marked with orange, blue and purple colour represent the miRNA, associated genes and associated TFs respectively)

H-test at 5% significance level. The same is also reported in Table 3 with the up/down regulation and PubMed ID. The *p-value* shows that the miRNAs are significantly differentially expressed with *p-value* less than 0.05 and PubMed ID are showing association

of these miRNAs to the breast cancer. Therefore, it is evident that the selected 10 miRNAs are quantitatively putative and statistically significant.

4.4 Biological Significance

The biological significance of the top 10 selected miRNAs is evaluated with the help of Protein-Protein Interaction (PPI) [25], KEGG pathway analysis [13] and Gene Ontology (GO) enrichment analysis [17] for Biological Process, Cellular Component and Molecular Function.

The Protein-Protein Interaction analysis for the selected miRNAs has been conducted using STRING [24] database and shown in Figure 2. In this diagram, each node represents protein produced by a single protein-coding gene locus, whereas each edge represents protein-protein associations. The degree of interaction of each node is computed and the top 10 proteins as transcription factors (TFs) are shown in Figure 2 with the help of a bar plot. It is observed from the analysis that some important human transcription factors (TFs) like *MYC*, *Estrogenreceptor1*, *HDAC2*, *CREB1* etc. for selected miRNAs are found as a part of Protein-Protein Interaction network. These TFs are found in breast cancer and can be regarded as targets for molecular therapies.

KEGG pathway analysis has been performed using DIANA tool [26] and the analytical findings uncover the pathway of targeted genes associated with the identified top 10 miRNAs. The targeted genes are extracted from miRTarBase database. Each pathway contains a particular score of adjusted *p-values* where, lower value signifies the higher probability of the pathway to be enriched with set of associated genes. Based on the values of adjusted *p-values*, top five pathways for selected top 10 miRNAs are shown in Figure 3. The analysis reveals the presence of PI3K-Akt signaling pathways which plays a significant role to stimulate the cell growth in human body. Over activation of this might cause an abnormal cell proliferation which are found at high rate in case of breast cancer [27]. We also found Proteoglycans which plays an important role in contributing to the various other cancer types. Similarly, FOXO signaling pathway is regarded as the target for the modulation of cancer [28].

Gene Ontology (GO) Enrichment analysis has been done using Enrichr tool [15]. This analysis discovers the various biological and cellular processes associated with selected 10 miRNAs as reported in Figures 4(a), (b) and (c). The biological process includes Cellular response to cytokine stimulus (GO:0071345), Cytokine-mediated signaling pathway (GO:0019221), Negative regulation of transcription, DNAtemplated (GO:0045892) etc. Similarly Cellular Components are found like Invadopodium (GO:0071437), Mitochondrial outer membrane (GO:0005741), Focal adhesion (GO:0005925), Cytoplasmic vesicle (GO:0031410) etc. and Molecular Functions like Ubiquitin protein ligase binding (GO:0031625), Protease binding (GO:0002020), Actin binding (GO:0003779) etc. are found as a part of GO enrichment analysis.

Additionally, Figure 5 shows the network analysis which has been performed using Cytoscape tool. The network analysis establishes the relationship among selected top 10 miRNAs with associated genes and transcription factors (TF). The associated genes and TFs are found as a part of KEGG pathway analysis and analysis of Protein-Protein Interaction. In the figure, the orange nodes represent the miRNA, whereas blue nodes signify associated genes and nodes with purple colour are associated TFs. From Figure 5, it is evident that hsa-mir-145 is associated with TF, *Estrogenreceptor1*

(*ESR1*) and Gene, *CYP2C19* which play a crucial role in breast cancer. Similarly, hsa-mir-182 is associated with *MYC*, *FOXO1* which also have important role to grow breast cancer.

5 CONCLUSION

In this article, a novel wrapper based feature selection technique has been proposed with the integration of clustering and classification tasks for selecting putative set of miRNAs. For this purpose, Rough and Fuzzy sets have been used to handle vagueness, uncertainty and overlapping characteristics of dataset while Random Forest and Particle Swarm Optimization have been used to improve the final results and to find the potential set of miRNAs by exploring the search space better. The results of the PSO-RFCM-RF have been demonstrated qualitatively and visually. It outperforms the existing techniques and provides putative miRNAs. Furthermore, the biological significance analysis has also been conducted to establish the biological relevance of those miRNAs in breast cancer. The results are statistically and biologically significant.

ACKNOWLEDGMENT

This work has been supported by Polish National Science Centre (2014/15/B/ST6/05082), Foundation for Polish Science (TEAM to DP) and the grant from Department of Science and Technology, Govt. of India and Polish Government under Indo-Polish/Polish-Indo project No.: DST/INT/POL/P-36/2016. The work was co-supported by grant 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation" within 4DNucleome NIH program, and by European Commission as European Cooperation in Science and Technology COST actions: CA18127 "International Nucleome Consortium" (INC), and CA16212 "Impact of Nuclear Domains On Gene Expression and Plant Traits". The work was partially supported as RENOIR Project by the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 691152 and by Ministry of Science and Higher Education (Poland), grant Nos. W34/H2020/2016, 329025/PnH/2016.

REFERENCES

- [1] N. S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.
- [2] C. H. Bang-Berthelsen, L. Pedersen, T. Fløyt, P. H. Hagedorn, T. Gylvin, and F. Pociot. 2011. Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. *BMC Genomics* 12, 1 (2011), 97.
- [3] D. P. Bartel. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* 136 (2009), 215–233.
- [4] S. Baskerville and D. P. Bartel. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* 11, 3 (2005), 241–247.
- [5] J. C. Bezdek. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic, MA, USA.
- [6] L. Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.
- [7] C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.
- [8] C. M. Croce. 2009. Causes and consequences of microRNA dysregulation in cancer. *Nature Reviews Genetics* 10 (2009), 704–714.
- [9] H. George and J. P. Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* 69 (1995), 338–345.
- [10] A. Grimson, K. K. Farh, W. K. Johnston, P. Garrett-Engle, L. P. Lim, and D. P. Bartel. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell* 27, 1 (2007), 91–105.

- [11] J. G. Hunsberger, E. B. Fessler, F. L. Chibane, Y. Leng, D. Maric, A. G. Elkahouloun, and D. M. Chuang. 2013. Mood stabilizer-regulated miRNAs in neuropsychiatric and neurodegenerative diseases: identifying associations and functions. *American Journal of Translational Research* 5, 4 (2013), 450–464.
- [12] A. Jacobsen, J. Wen, D. S. Marks, and A. Krogh. 2010. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome Research* 20 (2010), 1010–1019.
- [13] M. Kanehisa and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28 (2000), 27–30.
- [14] J. Kennedy and R. Eberhart. 1995. Particle swarm Optimization. In *Proceedings of IEEE International Conference on Neural Networks* 4 (1995), 1942–1948.
- [15] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* 44 (2016), W90–W97.
- [16] P. Maji and S. Paul. 2013. Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10, 2 (2013), 286–299.
- [17] A. Michael and et. al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25 (2000), 25–29.
- [18] Z. Pawlak. 1992. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, Norwell, MA, USA.
- [19] G. Peters, F. Crespo, P. Lingras, and R. Weber. 2013. Soft clustering - Fuzzy and rough approaches and their extensions and derivatives. *International Journal of Approximate Reasoning* 54, 2 (2013), 307–322.
- [20] J. R. Quinlan. 1986. Induction of Decision Trees. *Machine Learning* 1, 1 (1986), 81–106.
- [21] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Research* 14, 10A (2004), 1902–1910.
- [22] H. Song, Q. Wang, Y. Guo, S. Liu, R. Song, X. Gao, L. Dai, B. Li, D. Zhang, and J. Cheng. 2013. Microarray analysis of microRNA expression in peripheral blood mononuclear cells of critically ill patients with influenza A (H1N1). *BMC Infectious Diseases* 13, 1 (2013), 257.
- [23] Y. Sun, S. Koo, N. White, E. Peralta, C. Esau, N. M. Dean, and R. J. Perera. 2004. Development of a micro-array to detect human and mouse microRNAs and characterization of expression in human organs. *Nucleic Acids Research* 32 (2004), e188.
- [24] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian von Mering. 2017. The STRING database in 2017: quality-controlled protein – protein association networks, made broadly accessible. *Nucleic Acids Research* 45 (2017), D362–D368.
- [25] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. vonMering. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research* 45 (2017), D362–D368.
- [26] I. Vlachos, K. Zagkanas, M. D. Paraskevopoulou, G. Georgakilas, D. Karagkouni, T. Vergoulis, T. Dalamagas, and A. Hatzigeorgiou. 2015. DIANA-miRPath v3.0: Deciphering microRNA function with experimental support. *Nucleic Acids Research* 43 (2015), W460–W466.
- [27] S. X. Yang, E. Polley, and S. Lipkowitz. 2016. New insights on PI3K/AKT pathway alterations and clinical outcomes in breast cancer. *Cancer Treatment Review* 45 (2016), 87–96.
- [28] X. Zhang, N. Tang, T. J. Hadden, and A. Rishi. 2011. Akt, FoxO and regulation of apoptosis. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1813, 11 (2011), 1978–1986.



Identification of miRNA Biomarkers for Diverse Cancer Types Using Statistical Learning Methods at the Whole-Genome Scale

Jnanendra Prasad Sarkar^{1,2†}, Indrajit Saha^{3*†}, Adrian Lancucki^{4†}, Nimisha Ghosh⁵, Michal Wlasnowolski⁶, Grzegorz Bokota^{7,8}, Ashmita Dey², Piotr Lipinski⁴ and Dariusz Plewczynski^{6,8*}

OPEN ACCESS

Edited by:

Natalia Polouliakh,
Sony Computer Science Laboratories,
Japan

Reviewed by:

Jaromir Gumulec,
Masaryk University, Czechia
Tao Huang,
Shanghai Institute for Biological
Sciences (CAS), China

*Correspondence:

Indrajit Saha
indrajit@nittrkol.ac.in
Dariusz Plewczynski
d.plewczynski@cent.uw.edu.pl

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 25 October 2019

Accepted: 03 August 2020

Published: 13 November 2020

Citation:

Sarkar JP, Saha I, Lancucki A,
Ghosh N, Wlasnowolski M, Bokota G,
Dey A, Lipinski P and Plewczynski D
(2020) Identification of miRNA
Biomarkers for Diverse Cancer Types
Using Statistical Learning Methods at
the Whole-Genome Scale.
Front. Genet. 11:982.
doi: 10.3389/fgene.2020.00982

¹ Data, Analytics & AI, Larsen & Toubro Infotech Ltd., Pune, India, ² Department of Computer Science & Engineering, Jadavpur University, Kolkata, India, ³ Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India, ⁴ Computational Intelligence Research Group, Institute of Computer Science, University of Wrocław, Wrocław, Poland, ⁵ Department of Computer Science and Information Technology, SOA University, Bhubaneswar, India, ⁶ Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland, ⁷ Institute of Informatics, University of Warsaw, Warsaw, Poland, ⁸ Centre of New Technologies, University of Warsaw, Warsaw, Poland

Genome-wide analysis of miRNA molecules can reveal important information for understanding the biology of cancer. Typically, miRNAs are used as features in statistical learning methods in order to train learning models to predict cancer. This motivates us to propose a method that integrates clustering and classification techniques for diverse cancer types with survival analysis via regression to identify miRNAs that can potentially play a crucial role in the prediction of different types of tumors. Our method has two parts. The first part is a feature selection procedure, called the stochastic covariance evolutionary strategy with forward selection (SCES-FS), which is developed by integrating stochastic neighbor embedding (SNE), the covariance matrix adaptation evolutionary strategy (CMA-ES), and classifiers, with the primary objective of selecting biomarkers. SNE is used to reorder the features by performing an implicit clustering with highly correlated neighboring features. A subset of features is selected heuristically to perform multi-class classification for diverse cancer types. In the second part of our method, the most important features identified in the first part are used to perform survival analysis via Cox regression, primarily to examine the effectiveness of the selected features. For this purpose, we have analyzed next generation sequencing data from The Cancer Genome Atlas in form of miRNA expression of 1,707 samples of 10 different cancer types and 333 normal samples. The SCES-FS method is compared with well-known feature selection methods and it is found to perform better in multi-class classification for the 17 selected miRNAs, achieving an accuracy of 96%. Moreover, the biological significance of the selected miRNAs is demonstrated with the help of network analysis, expression analysis using hierarchical clustering, KEGG pathway analysis, GO enrichment analysis, and

protein-protein interaction analysis. Overall, the results indicate that the 17 selected miRNAs are associated with many key cancer regulators, such as MYC, VEGFA, AKT1, CDKN1A, RHOA, and PTEN, through their targets. Therefore the selected miRNAs can be regarded as putative biomarkers for 10 types of cancer.

Keywords: cancer, cox regression, feature selection, gene ontology, KEGG pathway, machine learning, next generation sequencing, stochastic neighbor embedding

1. INTRODUCTION

MicroRNAs (miRNAs) belong to the non-coding RNA family. They consist of 19–25 nucleotides and play an important role in the regulation of gene silencing. These non-coding RNAs are present in every eukaryotic cell and can also be encoded by a viral genome (Ray and Maiti, 2015; Bruscella et al., 2017). The miRNAs are formed by RNA polymerase II in the cell nucleus and are then transferred to the cytoplasm (Bartel, 2009) for biological activities such as cell cycle control, apoptosis, and oncogenesis. They interact with the complementary strand of mRNAs and lead to the degradation of the corresponding mRNAs; they also interfere with protein production by suppressing protein synthesis (Valencia-Sanchez et al., 2006). A miRNA molecule can bind one or more targets, thus forming a complex underlying regulatory network. These networks have a profound impact on cancer signaling pathways (Wang et al., 2017). Previously, low-throughput and high-cost technologies were the main obstacle to answering systems-level biological questions. However, recent advancements in next generation sequencing (NGS) have enabled researchers to address such complex problems (van Dijk et al., 2014). Moreover, new sequencing technologies and genomic datasets have helped us to gain better understanding of the biological complexities related to genomic abnormalities in cancer. The considerable achievements in sequencing techniques have made high-throughput techniques a fundamental platform for miRNA, RNA, and DNA research. Generally, miRNAs are involved in a wide range of diseases, including neurological disease, heart disease, and cancer (Giza et al., 2014; Paul et al., 2018). In many cases of cancer in humans, dysregulation of miRNA expression has been observed, and it is well-known that miRNAs can serve as potential cancer biomarkers (Lu et al., 2005; Jacobsen et al., 2013; Wong et al., 2017). In this regard, scientific communities are also trying to understand the role of miRNAs in paring with mRNAs (Zhang et al., 2014; Shrestha et al., 2017), in different cancer types by ranking miRNAs (Li et al., 2014) to elucidate their effects and drug resistance (Ma et al., 2010; Li and Yang, 2014; Cheerla and Gevaert, 2017).

To reduce the time taken for clinical trials, and to provide better and more accurate treatments while avoiding unnecessary interventions, the proper selection of miRNAs as biomarkers is crucial. For this purpose, miRNAs are often used as features in statistical learning methods viz. clustering, classification, and regression in order to identify potential biomarkers (Song et al., 2016; Yang et al., 2017; Yokoi et al., 2017). Song et al. (2016) performed a clustering analysis on breast cancer data in order to find miRNAs that could be prognostic biomarkers; these miRNAs

are up-regulated in this type of cancer and are linked to local relapse, distant metastasis, and poor clinical outcomes. Similarly, Yang et al. (2017) used clustering to find miRNA biomarkers for breast cancer. The identified miRNAs have higher specificity and sensitivity than single-gene biomarkers. On the other hand, Yokoi et al. (2017) used a classification task to inform the development of a predictive model to distinguish patients with ovarian cancer tumors from healthy subjects. In this study, eight miRNAs were found as biomarkers for ovarian cancer. Jacob et al. (2017) conducted a study on colon cancer and identified 16 miRNA signatures, which act as prognostic biomarkers at cancer stages II and III. Apart from the aforementioned works, regression analysis has been used to predict the survival rates of patients with different types of cancer (Liang et al., 2018). Liang et al. (2018) used Cox regression analysis for pancreatic cancer and identified five miRNAs as independent prognostic factors. The progress in this regard can be found in the literature (e.g., Peng and Croce, 2016; Hosseini et al., 2018).

Generally, in miRNA-based cancer studies, statistical learning algorithms viz. clustering, classification, and regression are used separately for different cancer types, as described above and in the literature (Akhtar et al., 2015; Ang et al., 2016; Li et al., 2016; Lin and Lane, 2017). However, to leverage the advantages of the different algorithms, it may be useful to integrate them into a single method for identifying potential biomarker miRNAs. Besides, little work exists on multi-class classification of diverse cancer types using NGS data. These two facts motivated us to develop the method described in this paper, which can not only classify 10 types of cancer (bladder, breast, colon, glioblastoma, head and neck squamous cell, kidney renal clear cell, lung adenocarcinoma, lung squamous cell, ovarian, and uterine corpus endometrial carcinoma) but also find putative miRNAs that are highly associated with these cancer types. The proposed two-part wrapper-based feature selection method, referred to as the stochastic covariance evolutionary strategy with forward selection (SCES-FS), uses stochastic neighbor embedding (SNE) (Hinton and Roweis, 2003) in conjunction with the covariance matrix adaptation evolutionary strategy (CMA-ES) (Hansen et al., 2003) and a simple classifier, either random forest (RF) (Breiman, 2001), support vector machine (SVM) (Cortes and Vapnik, 1995), naive Bayes (NB) classifier (George and Langley, 1995), *K*-nearest-neighbors (*K*-NN) classifier (Altman, 1992), or decision tree (DT) (Quinlan, 1986), in the first part. Here SNE is used to reorder the features by performing an implicit clustering, such that neighboring features are highly correlated. Then, from these clusters of features, a subset of features is selected randomly in order to perform the multi-class

classification task for the diverse cancer types. However, as the features are randomly selected, this classification task is treated as an underlying optimization problem for CMA-ES to find the features automatically. Hence, the final set of features/miRNAs is obtained by using forward selection (Whitney, 1971).

In the second part of the method, survival analysis is performed using Cox regression (Cox, 1972), where the levels of miRNA expression and corresponding clinical data are used. The experiment is conducted with data collected from The Cancer Genome Atlas (TCGA)¹ for 10 different types of cancer. The performance of the proposed wrapper-based feature selection method is compared with the following methods in terms of classification accuracy, with the top 17 miRNAs selected as putative biomarkers: ensemble SVM-recursive feature elimination (ESVM-RFE) (Anaissi et al., 2016), the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), the non-dominated sorting genetic algorithm II-based stacked ensemble (NSGA-II-SE) (Saha et al., 2017), the SVM-wrapped multi-objective genetic algorithm (MOGA) (Mukhopadhyay and Maulik, 2013), SVM-based novel recursive feature elimination (SVM-nRFE) (Peng et al., 2009), SVM recursive feature elimination (SVM-RFE) (Guyon et al., 2002), conditional mutual information (CMIM) (Fleuret, 2004), interaction capping (ICAP) (Jakulin, 2005), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), joint mutual information (JMI) (Bennasar et al., 2015), conditional infomax feature extraction (CIFE) (Brown et al., 2012), minimum redundancy maximum relevance (mRMR) (Peng et al., 2005), feature selection with Cox regression (FSCOX) (Kim et al., 2014), double-input symmetrical relevance (DISR) (Brown et al., 2012), signal-to-noise ratios (SNRs) (Mishra and Sahu, 2011), and the Wilcoxon rank-sum test (RankSum) (Troyanskaya et al., 2002). Thereafter, the significance of the 17 selected miRNAs to 10 different cancer types is determined using Cox regression analysis. Finally, survival analysis, network analysis, expression analysis using hierarchical clustering in the form of heatmaps, KEGG pathway analysis (Kanehisa and Goto, 2000), gene ontology (GO) enrichment analysis (Kuleshov et al., 2016), and protein-protein interaction (PPI) network analysis (Szklarczyk et al., 2019) are performed to assess the biological significance of the selected miRNAs. Additionally, a web-based cancer predictor application is developed to predict 10 different types of cancer given the expression of 17 miRNAs.

2. MATERIALS AND METHODS

In this section, we briefly describe SNE (Hinton and Roweis, 2003), CMA-ES (Hansen, 2006), and Cox regression analysis (Cox, 1972). More details about classification techniques and feature selection methods are given in the **Supplementary Material**² for this article. This section also describes the proposed method, which consists of two parts. The first part is the wrapper-based SCES-FS. In the second part, the selected features, such as expression of miRNAs and clinical

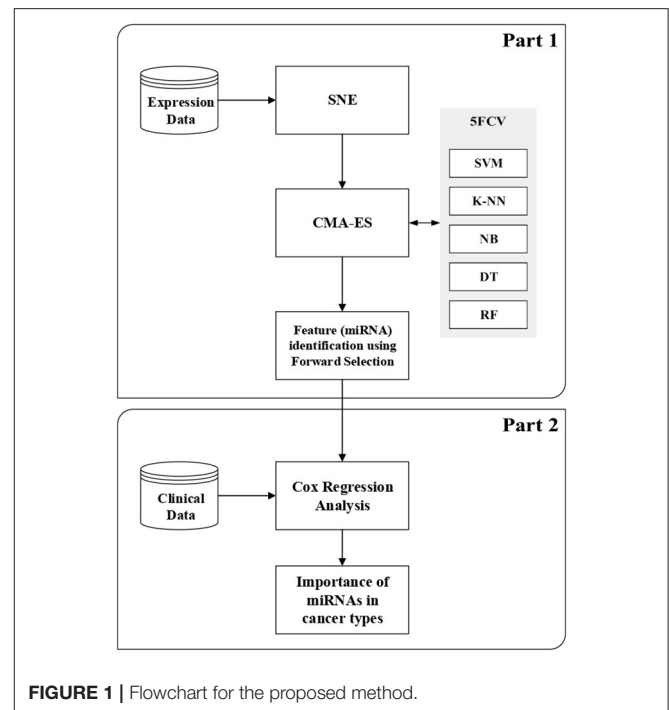


FIGURE 1 | Flowchart for the proposed method.

data, are used in survival analysis to assess the importance of the selected miRNAs in different types of cancer and to evaluate the effectiveness of the selected miRNAs. **Figure 1** shows the flowchart of the proposed method.

2.1. Stochastic Neighbor Embedding

Let $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ denote a set of n observations, where $x_i \in \mathbb{R}^D$. SNE (Hinton and Roweis, 2003) constructs a low-dimensional embedding that recreates \mathcal{X} in a space of lower dimension as $\mathcal{X}' = \{x'_1, x'_2, \dots, x'_n\}$, where $x'_i \in \mathbb{R}^d$. In SNE, both \mathcal{X} and \mathcal{X}' are represented as discrete probability distributions P and Q , where

$$p_{ij} = \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2 \text{var}_i}\right)}{\sum_{k \neq i} \exp\left(\frac{-\|x_i - x_k\|^2}{2 \text{var}_i}\right)}, \quad q_{ij} = \frac{\exp\left(-\|x'_i - x'_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|x'_i - x'_k\|^2\right)}, \quad (1)$$

that model pairwise distances between data points. The values of $\text{var}_i \in \mathbb{R}$ are adjusted in such a way that the entropies of all distributions P_i are equal.

The mismatch between P and Q is reduced through minimization of Kullback-Leibler (KL) divergence objective $KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} = \sum_i KL(P_i||Q_i)$, by altering Q with gradient-based optimization methods. Optimization is difficult due to the existence of multiple local optima, and entirely different embeddings may be obtained with different initial Q distributions.

¹<https://tcga-data.nci.nih.gov/tcga/>

²<http://www.nitttrkol.ac.in/indrajit/projects/mirna-prediction-multiclass/>

2.2. Covariance Matrix Adaptation Evolution Strategy

Evolutionary strategies are black-box optimization algorithms which belong to a broader group of evolutionary algorithms. In such methods, a set of candidate solutions is maintained. In successive iterations of the procedure, these candidate solutions are perturbed and evaluated, and in each iteration the best solution is left unchanged and carried over to the next set of candidate solutions. CMA-ES (Hansen and Ostermeier, 1996) is an evolutionary strategy where the set of candidate solutions is modeled and sampled from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{m}, \mathbf{C})$. The covariance matrix \mathbf{C} represents pairwise relationships between attributes. The objective function of CMA-ES maximizes two likelihoods: (a) the likelihood of having the best individuals in previous iterations, and (b) the likelihood of taking the best search steps in previous iterations. At the end of each iteration, (a) guides updates of the mean \mathbf{m} as a weighted average of μ best solutions, $\mathbf{m}^{(g+1)} = \sum_{i=0}^{\mu} w_i x_i^{(g+1)}$, where $x_i^{(g+1)}$ is the i th best solution in iteration $g+1$ and w_i is its weight, and (b) updates the covariance matrix \mathbf{C} as follows:

$$\mathbf{C}^{(g+1)} = (1 - c_1 - c_\mu) \mathbf{C}^{(g)} + c_1 \mathbf{p}_c^{(g+1)} \mathbf{p}_c^{(g+1)T} + c_\mu \sum_{i=1}^{\mu} w_i \left(\frac{x_i^{(g+1)} - \mathbf{m}^{(g)}}{\sigma^{(g)}} \right) \left(\frac{x_i^{(g+1)} - \mathbf{m}^{(g)}}{\sigma^{(g)}} \right)^T \quad (2)$$

Finally, $\mathbf{p}_c^{(g+1)} \in \mathbb{R}^D$ is a vector that amplifies the updates in favorable directions:

$$\mathbf{p}_c^{(g+1)} = (1 - c_c) \mathbf{p}_c^{(g)} + z \frac{\mathbf{m}^{(g+1)} - \mathbf{m}^{(g)}}{\sigma^{(g)}}, \quad (3)$$

where $c_1, c_\mu, c_c \in \mathbb{R}$ are weights, $\sigma^{(g)} \in \mathbb{R}$ is an adaptive step size, which is dependent on the iteration, and $z \in \mathbb{R}$ is a normalizing constant. Details of the parameters of CMA-ES can be found in Hansen and Ostermeier (1996) and Hansen et al. (2003).

2.3. Cox Regression Analysis

The Cox regression model (Cox, 1972) is a proportional hazards regression model in which the hazard ratio is constant but other contents have the same baseline hazard function. Based on this assumption, the survival function is calculated as

$$\mathbb{S}(\tau) = \exp(-H_0(\tau) \exp(X\beta)) = \mathbb{S}_0(\tau)^{\exp(X\beta)}, \quad (4)$$

where $H_0(\tau)$ represents the cumulative baseline hazard function at time τ and $\mathbb{S}_0(\tau) = \exp(-H_0(\tau))$ is the baseline survival function; $H_0(\tau)$ is taken to be Breslow's estimator (Breslow, 1974), which is the most widely used and given by

$$\hat{H}_0(\tau) = \sum_{\tau_i \leq \tau} \hat{h}_0(\tau_i). \quad (5)$$

As the Cox model is based on the proportional hazards assumption, it is represented as

$$h(\tau, x_i) = h_0(\tau) \exp(x_i \beta) \quad (6)$$

for an given instance $i = 1, 2, 3, \dots, n$, where the baseline hazard function $h_0(\tau)$ can be an arbitrary negative function of time, and $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$ is the corresponding covariate vector for instance i and is the coefficient vector. The Cox model is a semi-parametric algorithm where the baseline hazard function $h_0(\tau)$ is unspecified. For any two instances x_1 and x_2 , the hazard ratio is given by

$$\frac{h(\tau, x_1)}{h(\tau, x_2)} = \frac{h_0(\tau) \exp(x_1 \beta)}{h_0(\tau) \exp(x_2 \beta)} = \exp[(x_1 - x_2) \beta]. \quad (7)$$

This means that the hazard ratio is independent of the baseline hazard function.

2.4. Wrapper-Based Feature Selection Integrating SNE and CMA-ES

The first part of our proposed method performs the task of miRNA selection for diverse cancer types, which is considered a multi-class classification problem here. SNE is used to reorder the features by performing an implicit clustering such that neighboring features are highly correlated. Then, the underlying multi-class classification task is performed using well-known classifiers and treated as an optimization problem for which CMA-ES is used to find the miRNAs automatically. The miRNAs thus found are further refined using forward selection (FS). Therefore, we call this wrapper-based feature selection method as stochastic covariance evolutionary strategy with forward selection (SCES-FS).

Algorithm 1 presents the SCES-FS method in detail. It starts with the dataset $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$, which denotes a set of n observations with $x_i \in \mathbb{R}^D$, the class label \mathcal{Y} , the population size λ , the maximum number of generations N , the classifier \mathcal{A} , and the number of runs \mathcal{R} as inputs. In the dataset, each feature is characterized by expression levels of the samples. The features of the original dataset \mathcal{X} are reordered using SNE in the *ConstructEmbedding* step, producing the embedding dataset \mathcal{X}' whose size is the same as that of the original dataset. The parameters are initialized in the *SetParamsCMAES* step. The individuals/vectors in CMA-ES are encoded as a simple threshold weight vector, with a single weight for each miRNA. An individual $x \in \mathbb{R}^H$ encodes a weight vector $w \in \mathbb{R}^H$ and a threshold $t \in \mathbb{R}$, where $H \leq D$ is the number of weights. Only those features whose weights exceed the threshold are eventually selected into a feature set S :

$$S = \{i \in 1, \dots, D : \text{closest}(x_i) = j \wedge w_i \geq t\}, \quad (8)$$

where $t \in \mathbb{R}$ is a threshold that is chosen carefully. The population of vectors is drawn in the *DrawPopulationCMAES* step from $\mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{C})$, where \mathbf{m} is the mean, \mathbf{C} is the covariance matrix, and σ is the step-size control parameter of CMA-ES. In the *ConstructFeatureSets* step, each individual x_i is translated to a feature set S_i according to Equation (8).

In the *ScoreFeatureSet* step, each feature set is evaluated by training a classifier on a dataset restricted to only the selected features. The objective function of CMA-ES combines

the accuracy obtained from the classifier for each individual x and the size of the feature set S as follows:

$$f(x) = \text{accuracy}(S) - \alpha \frac{\max(0, |S| - \Upsilon)}{D}, \quad (9)$$

where D is the dimension of the dataset, Υ is the target number of features, and $\alpha \in \mathbb{R}_+$ balances the penalty for the excessive number of features. The objective function is designed in this way so that higher classification accuracy can be achieved for a small number of features, and it incorporates L_0 regularization term as penalty, which is bounded by Υ in order to have redundancy in selected subsets of features. The optimization with CMA-ES allows inter-feature relationships with covariance matrices. Finally, the parameters are updated according to CMA-ES update rules in the *UpdateParamsCMAES* step, and the best set of features/miRNAs in a particular run are kept in S_{RBest} .

Algorithm 1 Pseudo-code of the SCES-FS

Input: $\mathcal{X}, \mathcal{Y}, \lambda, N, \mathcal{A}, \mathcal{R}$ (dataset, class label, population size, maximum number of generations, classifier, number of runs)

Output: S_{best} (feature subset)

```

1: Initialize a NULL list,  $L$ 
2: for  $i \leftarrow 1$  to  $\mathcal{R}$  do
3:    $\mathcal{X}' \leftarrow \text{ConstructEmbedding}(\mathcal{X})$  // With SNE
4:    $\theta \leftarrow \text{SetParamsCMAES}()$  //  $\mu, \sigma, \mathbf{m}, \mathbf{C}$ 
5:   for  $g \leftarrow 1$  to  $N$  do
6:      $\mathcal{P} \leftarrow \text{DrawPopulationCMAES}(\mathcal{X}', \lambda, \theta)$ 
7:      $S \leftarrow \text{ConstructFeatureSets}(\mathcal{P})$ 
8:      $\text{ScoreFeatureSet}(S, \mathcal{X}', \mathcal{A}, \mathcal{Y})$ 
9:      $S_{\text{RBest}} \leftarrow \text{BestFeatureSet}(S, S_{\text{RBest}})$ 
10:     $\theta \leftarrow \text{UpdateParamsCMAES}(\theta)$ 
11:   end for
12:    $L \leftarrow L \cup S_{\text{RBest}}$ 
13: end for
14:  $L' \leftarrow \text{RankFrequency}(L)$ 
15:  $S_{\text{best}} \leftarrow \text{ForwardSelection}(L', \mathcal{X}', \mathcal{A}, \mathcal{Y})$ 
16: return  $S_{\text{best}}$ 

```

2.5. Preparation of the Final Set of Features

The SCES-FS is random in nature to reduce the probability of returning sub-optimal solutions. Therefore, a single run of SCES-FS does not guarantee a reliable solution. To overcome this limitation, SCES-FS is executed up to a maximum number of runs, \mathcal{R} . In each run, a set of best features/miRNAs are collected into a list, L . After completion of the maximum number of runs, *RankFrequency* sorts the cumulative set of features in descending order according to the frequency of occurrence in each run and produces a modified list, L' . Thereafter, *ForwardSelection* applies the forward feature selection method, using the classifier to evaluate the feature set iteratively to obtain the best feature set, S_{best} .

2.6. Justification for miRNA Selection

Because of the random nature of the algorithm, there is a chance of false positives or false negatives occurring in the selection

of miRNAs. To reduce the probability of having false positives or false negatives in the selected set of miRNAs, the SCES-FS algorithm is run 50 times, and then the miRNAs are ranked based on their frequencies of occurrence in 50 different sets of features. Thus, a stable set of miRNAs is selected on the basis of maximum classification accuracy, which is computed by considering the miRNAs cumulatively from the top of the list. By this process, 17 distinct miRNAs are selected.

This procedure does not, however, ensure the absence of false positives or false negatives, so additional measures are taken. The presence of false positives is made unlikely by the sorting procedure. In fact, since a given false positive is supposed to occur less frequently than all the true positives, it will appear in the tail after sorting. Consequently, it is likely that this false positive will be filtered out when selecting the final list of miRNAs. On the other hand, the presence of false negatives is related to the choice of the number of runs and the corresponding expected number of miRNAs belonging to the sorted list. In this regard, a mathematical argument is given below to justify that the SCES-FS does not exhibit random behavior.

Suppose that at each run the SCES-FS randomly selects 10 miRNAs from the whole collection of miRNAs. We first compute the expected number of distinct miRNAs reported after all the runs, and then we compare this number with the results of our experiments. We have the following parameters: $D = 199$ is the number of miRNAs, $S = 10$ is the number of miRNAs returned after each run, and $\mathcal{R} = 50$ is the number of runs. For $i = 1, \dots, D$, let V_i be a Bernoulli-distributed indicator variable, where $V_i = 1$ if the miRNA m_i never shows up. The probability that m_i is selected in one run is S/D , so the probability that it is never selected is $1 - S/D$. Since each of the \mathcal{R} runs is independent, the following equation can be written:

$$\mathbb{E}[V_i] = \Pr(V_i = 1) = \left(1 - \frac{S}{D}\right)^{\mathcal{R}}. \quad (10)$$

Let $V = \sum_{i=1}^D V_i$ be the random variable that counts the number of miRNAs that do not belong to the final set of miRNAs reported at least once. By linearity of the expectation, we obtain the equation

$$\mathbb{E}[V] = \sum_{i=1}^D \mathbb{E}[V_i] = D \left(1 - \frac{S}{D}\right)^{\mathcal{R}}. \quad (11)$$

Hence, the number of expected miRNAs reported at least once is

$$\mathbb{E}[(D - V)] = D - \mathbb{E}[V]. \quad (12)$$

Substituting the above-mentioned values for the parameters, we obtain that

- the expected number of miRNAs reported at least once after 50 iterations is 183; and
- the number of new miRNAs added in a further iteration would be 1.

As the number of sorted miRNAs in our experiment is 39 (see the **Supplementary Material**), this difference suggests that 50 runs

TABLE 1 | Details of the data for 10 different cancer types.

Cancer type	Code	No. of tumor samples	Gender		Average age	Average no. days from last followup
			Male	Female		
Bladder urothelial carcinoma	BLCA	94	67	27	67.07	416.97
Breast invasive carcinoma	BRCA	255	0	255	58.28	1297.82
Colon adenocarcinoma	COAD	119	57	62	70.91	616.80
Glioblastoma multiforme	GBM	38	20	18	62.78	376.81
Head and neck squamous cell carcinoma	HNSC	298	218	80	60.98	1038.31
Kidney renal clear cell carcinoma	KIRC	146	91	55	59.99	1236.55
Lung adenocarcinoma	LUAD	61	29	32	65.62	743.09
Lung squamous cell carcinoma	LUSC	89	64	25	64.78	1296.87
Ovarian serous cystadenocarcinoma	OV	509	0	509	59.85	1020.87
Uterine corpus endometrial carcinoma	UCEC	98	0	98	62.32	1066.01

TABLE 2 | Values of parameters.

Symbol	Value	Description
C	I	CMA-ES initial covariance matrix of size H
σ	0.3	Initial value of step-size control parameter
λ	200	Population size
μ	100	Number of parents
\mathcal{R}	50	Number of runs
N	200	Maximum number of generations
t	0.5	Threshold for calculating subsets S_i
α	0.5	Excessive attributes penalty term
Υ	10	Target number of miRNAs
γ	0.05	SVM RBF kernel parameter
C	1.0	SVM C constant
K	5	Value of K in K -NN
\mathcal{M}	50	Number of trees in RF

are enough to conclude that all the true positives are included in the sorted list and so false negatives are unlikely.

2.7. Cox Regression Analysis for Evaluating miRNAs

The second part of the proposed method is for the evaluation of selected miRNAs using Cox regression analysis. The primary objective of this stage is to assess the importance of the miRNAs selected in the first part of the method, and this is done using Cox regression analysis for survival in 10 different cancer types. Expression data of the selected miRNAs and the associated clinical data are used for the Cox regression analysis. In the clinical data, the vital status of each patient, indicating whether the patient is still alive or has died, and the number of days since the last followup are taken into account when performing the survival analysis. Based on the expression levels and the clinical data of the selected miRNAs in each cancer type, the Cox coefficient, hazard ratio, and p -value are computed. A higher value of the Cox coefficient signifies greater importance of that miRNA to the respective cancer type. Moreover, the up- and

down-regulation of all the selected miRNAs are observed to understand their behavior with respect to that particular cancer type based on change in expression in tumor and normal samples.

2.8. Complexity Analysis

Let D be the number of features and n the number of samples in the input dataset. As the available approximations may considerably lower the overall complexity, we discuss the complexity of each building block separately. A single step of SNE requires computing relations between all data points. We embed a transposed dataset, making an optimization step in $O(D^2)$ time. Computing SNE usually involves performing principal component analysis for preliminary dimension reduction, though its cost is negligible. The internal complexity of CMA-ES is estimated as $O(D^2)$, due to sampling and updating of the covariance matrix. The matrix needs to be factorized, which can be done by eigen decomposition in $O(D^3)$ time. Factorization does not happen in every generation, which gives $O(D^2)$ amortized time. Empirical evidence suggests that the sufficient number of objective function evaluations usually scales sub-quadratically with D (Ros and Hansen, 2008). The computation time is similar, since the vast majority of it is time spent training similar classifiers. On the other hand, the time complexity of Cox regression analysis is $O(nD^2)$ for a single run (Kelley, 1999).

3. RESULTS AND DISCUSSION

The performance of the proposed method (SCES-FS) was tested on real miRNA expression datasets for 10 different cancer types and compared with the results of 16 existing methods, namely ESVM-RFE (Anaissi et al., 2016), LASSO (Tibshirani, 1996), NSGA-II-SE (Saha et al., 2017), MOGA (Mukhopadhyay and Maulik, 2013), SVM-nRFE (Peng et al., 2009), SVM-RFE (Guyon et al., 2002), CMIM (Fleuret, 2004), ICAP (Jakulin, 2005), SCAD (Fan and Li, 2001), JMI (Bennasar et al., 2015), CIFE (Brown et al., 2012), mRMR (Peng et al., 2005), FSCOX (Kim et al., 2014), DISR (Brown et al., 2012), SNRs (Mishra and Sahu, 2011), and RankSum (Troyanskaya

TABLE 3 | Number of features and classification accuracy of feature selection methods for five classifiers with five-fold cross-validation.

Method	Number of features	RF	SVM	NB	K-NN	DT
SCES-FS	17	96.881 ± 0.039	96.332 ± 0.194	96.251 ± 0.168	96.132 ± 0.369	94.232 ± 0.057
ESVM-RFE	22	95.684 ± 0.031	95.902 ± 0.193	92.672 ± 0.161	91.382 ± 0.369	90.429 ± 0.056
LASSO	48	95.601 ± 0.038	95.547 ± 0.191	92.582 ± 0.164	91.251 ± 0.367	90.241 ± 0.051
NSGA-II-SE	26	95.587 ± 0.033	95.537 ± 0.195	92.538 ± 0.166	91.183 ± 0.366	90.229 ± 0.052
MOGA	24	95.391 ± 0.036	95.531 ± 0.194	92.293 ± 0.161	90.338 ± 0.362	89.142 ± 0.055
SVM-nRFE	26	95.224 ± 0.032	95.321 ± 0.191	92.281 ± 0.167	90.106 ± 0.361	88.993 ± 0.051
SVM-RFE	28	95.048 ± 0.038	95.159 ± 0.199	92.116 ± 0.165	89.889 ± 0.366	88.691 ± 0.053
CMIM	28	94.299 ± 0.037	92.029 ± 0.193	90.683 ± 0.161	89.374 ± 0.369	89.161 ± 0.052
ICAP	27	93.721 ± 0.031	92.951 ± 0.192	90.874 ± 0.163	90.643 ± 0.362	87.057 ± 0.053
SCAD	25	91.972 ± 0.034	90.839 ± 0.194	90.003 ± 0.165	89.918 ± 0.366	87.495 ± 0.058
JMI	28	91.718 ± 0.031	90.602 ± 0.196	89.986 ± 0.166	88.639 ± 0.369	87.057 ± 0.051
CIFE	32	90.886 ± 0.034	89.072 ± 0.199	88.389 ± 0.162	87.261 ± 0.362	86.205 ± 0.056
mRMR	28	91.063 ± 0.032	89.753 ± 0.195	87.402 ± 0.161	87.254 ± 0.361	85.208 ± 0.059
FSCOX	23	89.298 ± 0.038	88.529 ± 0.198	87.505 ± 0.169	87.287 ± 0.368	85.498 ± 0.058
DISR	29	89.286 ± 0.031	88.276 ± 0.196	87.580 ± 0.167	87.321 ± 0.369	85.858 ± 0.053
SNR	30	87.866 ± 0.035	86.712 ± 0.193	85.749 ± 0.163	85.364 ± 0.365	84.042 ± 0.059
RankSum	32	86.633 ± 0.033	85.556 ± 0.199	85.466 ± 0.166	84.669 ± 0.366	84.322 ± 0.055
Without feature selection	199	86.428 ± 0.036	85.294 ± 0.191	85.183 ± 0.162	84.552 ± 0.367	84.118 ± 0.053

et al., 2002) (see section 1 for the full names of these methods), as well as the results with all features (i.e., without feature selection).

3.1. Dataset Preparation and Parameter Setting

The miRNA expression and clinical datasets of bladder, breast, colon, glioblastoma, head and neck squamous cell, kidney renal clear cell, lung adenocarcinoma, lung squamous cell, ovarian, and uterine corpus endometrial carcinoma were obtained from TCGA. These 10 cancer types have also been studied previously (see, e.g., Jacobsen et al., 2013). Moreover, Hoadley et al. (2018) found that the characteristics of certain cancers out of 33 types provided in TCGA are overlapping in nature. As a result, 10–15 distinct groups of cancer were reported in Hoadley et al. (2018), which are similar to the cancer types studied in the present article. Our choice of cancer types was based on: (1) careful review of the literature, (2) the availability of tissue-specific tumor and normal samples to avoid the class imbalance problem in classification, and (3) the availability of common miRNA expression data for different cancer types and their corresponding clinical data. Thus, 10 cancer types were selected for our study. The expression data were generated using an Illumina high-throughput sequencing machine in the form of read counts of 199 miRNAs, normalized to reads per million, while the clinical data contain *gender*, *age*, *days since last followup*, and *vital status*. After removing miRNAs that contain more than 60% zeros in each cancer type and taking the miRNAs common to all the cancer types, we found 199 miRNAs and considered their expression in different cancer types. The number of samples and other details for each cancer type are given in **Table 1**; we also included 333 normal samples in the analysis with the expression of same miRNAs. For

convenience, the expression datasets containing reads per million are further normalized onto a \log_2 scale. These processed datasets can be downloaded from the **Supplementary Material** website³. To construct the final ranking of miRNAs, the SCES-FS algorithm was run 50 times. Five-fold cross-validation was applied during each classification to avoid the issue of overfitting or underfitting. The parameters used in the experiments are shown in **Table 2** and were obtained either experimentally or from the literature (Latinne et al., 2001; Oshiro et al., 2012; Hansen, 2016).

3.2. Experimental Outcomes

The problem of finding miRNAs that can correctly distinguish different cancer types were posed as a multi-class classification task using expression data of miRNAs, and the importance of the selected miRNAs was evaluated using Cox regression analysis with the help of clinical data. The classification results of SCES-FS using the classifiers RF, SVM, NB, K-NN, and DT over 50 runs are reported in **Table 3** and compared with the results of ESVM-RFE, LASSO, NSGA-II-SE, MOGA, SVM-nRFE, SVM-RFE, CMIM, ICAP, SCAD, JMI, CIFE, mRMR, FSCOX, DISR, SNR, and RankSum, as well as the results with all features (i.e., without feature selection). These results show that SCES-FS with RF achieved the highest classification accuracy, 96.881, with a standard deviation of 0.039, while the accuracy results of SCES-FS with SVM, NB, K-NN, and DT are 96.332 ± 0.194 , 96.251 ± 0.168 , 96.132 ± 0.369 , and 94.232 ± 0.057 , respectively, still better than those of the other existing feature selection methods.

It is to be noted that our results have been verified with FSCOX, which computes important miRNAs via Cox regression on all miRNAs and subsequently uses

³<http://www.nitttrkol.ac.in/indrajit/projects/mirna-prediction-multiclass/>

them for the classification. It has been observed that the overall classification accuracy of FSCOX is less than the accuracy attained by SCES-FS as reported in **Table 3**. Other omics-based analyses, such as protein array, copy number variation (CNV), and methylation studies, also have potential applications in pan-cancer classification, as explained in Zhang et al. (2015, 2016, 2019), where the overall accuracy achieved in the range of 93–97%. The cancer dataset used in Zhang et al. (2015, 2016, 2019) was carefully selected for protein array, CNV, and methylation studies and is not directly suitable for experimenting with miRNAs, as it creates a class imbalance problem when selecting miRNA expression data for both tumor and normal cases. However, the overall accuracy of our method is higher than 96%, which is on the higher side of the accuracy range reported by Zhang et al. (2015, 2016, 2019), suggesting that our selected miRNAs can also be considered potential markers for pan-cancer classification.

The results of SCES-FS have been further validated by survival analysis, including Cox regression, network analysis, expression analysis using hierarchical clustering in the form of heatmaps, KEGG pathway analysis, GO enrichment analysis, and PPI network analysis, as described in the following.

3.2.1. Survival Analysis

Table 4 reports the results of the Cox regression analysis, i.e., the Cox coefficient and hazard ratio values, for 10 cancer types in order to evaluate the importance of the 17 selected miRNAs with respect to these cancer types. The Cox regression analysis was performed by integrating the miRNA expression and clinical data to assess the effect of miRNA expression on cancer survival. Higher Cox coefficient and hazard ratio values indicate greater influence of the miRNA on the cancer type. For example, the miRNA hsa-mir-375 has the highest Cox coefficient, 0.9882, and hazard ratio, 1.7879, for the GBM cancer type. To help visualize the importance of each miRNA to the different cancer types, a circos plot of **Table 4** is shown in **Figure 2**. In the figure, a broader band signifies a stronger association between the miRNA and the particular cancer type. **Table 5** summarizes the cancer types that are most highly associated with the selected miRNAs. In this table, for each miRNA, the cancer type for which that miRNA has the highest Cox coefficient is shown, along with the associated *p*-value, false discovery rate (FDR), and up- or down-regulation of the miRNA expression in that cancer type. Similarly, **Supplementary Table 2** gives the up- and down-regulations along with the corresponding FDR values of the selected miRNAs for all cancer types. The up- and down-regulations of the miRNAs are computed after evaluating the change in expression of the selected miRNAs between tumor samples and normal samples or the population for a particular cancer type using the ANOVA test; this is discussed further in section 3.2.3. Here, the change in expression is considered significant if the *p* < 0.05, and the up- or down-regulation is computed if the change in population expression is positive or negative, respectively. In summary, we find that hsa-mir-205, hsa-mir-10a, hsa-mir-196b, hsa-mir-10b, hsa-mir-375, hsa-mir-143, hsa-let-7c, hsa-mir-107, hsa-mir-378, hsa-mir-133a, hsa-mir-1, hsa-mir-30c,

TABLE 4 | Results of Cox regression analysis of each selected miRNA.

miRNA	BLCA		BRCA		COAD		GBM		HNSC		KIRC		LUAD		LUSC		OV		UCEC	
	Cox	Hazard	Cox	Hazard	Cox	Hazard	Cox	Hazard	Cox	Hazard	Cox	Hazard	Cox	Hazard	Cox	Hazard	Cox	Hazard	Cox	Hazard
hsa-mir-205	-0.0344	0.9661	-0.0795	0.9234	0.1109	1.1172	0.9233	2.5177	0.0586	1.0600	0.0187	1.0189	0.1292	1.1379	-0.0449	0.9560	-0.0224	0.9778	-0.0039	0.9980
hsa-mir-10a	0.0618	1.0637	-0.0750	0.9277	-0.1644	0.8483	0.4881	1.6293	-0.0428	0.9580	-0.0122	0.9878	0.1193	1.1267	0.1619	1.1758	0.0569	1.0586	-0.4075	0.6652
hsa-mir-196b	-0.0333	0.9672	0.0089	1.0090	-0.2236	0.7996	0.2573	1.2935	-0.0095	0.9905	0.1367	1.1465	0.1124	1.1189	-0.0282	0.9721	0.0146	1.0147	0.4581	1.5811
hsa-mir-10b	0.0828	1.0864	-0.2213	0.8014	0.4042	1.4981	0.0795	1.0827	0.0597	1.0615	-0.2743	0.7600	0.1626	1.1766	0.0176	1.0178	-0.0218	0.9783	0.5080	1.6621
hsa-mir-375	-0.0226	0.9776	0.1334	1.1427	-0.0313	0.9691	0.9882	1.7879	0.0525	1.0539	-0.0051	0.9948	-0.0967	0.9078	-0.0340	0.9665	-0.0768	0.9260	0.0243	1.0246
hsa-mir-143	0.1158	1.1228	-0.2472	0.7809	-0.0759	0.9268	-0.0654	0.9366	-0.0630	0.9389	-0.0433	0.9575	-0.1486	0.8619	0.0053	1.0054	0.1504	1.1623	0.0770	1.0801
hsa-let-7c	0.1731	1.1890	-0.1825	0.8331	0.0702	1.0727	-0.1349	0.8737	0.0430	1.0440	-0.1158	0.8905	-0.1544	0.8569	0.0665	1.0687	0.0618	1.0637	0.0090	1.0090
hsa-mir-107	-0.0395	0.9611	0.1127	1.1193	0.0448	1.0459	0.5363	1.7096	-0.0389	0.9617	0.0405	1.0413	0.0885	1.0925	-0.0926	0.9114	-0.0919	0.9121	0.0609	1.0628
hsa-mir-378	-0.0286	0.9717	-0.1355	0.8732	0.1520	1.1641	0.5511	1.7352	0.0688	1.0712	-0.1881	0.8284	-0.1228	0.8843	0.1748	1.1910	0.0566	1.0583	-0.1596	0.8524
hsa-mir-133a	0.0610	1.0629	0.1604	1.1740	-0.1025	0.9025	0.2683	1.4481	-0.0393	0.9614	0.0328	1.0334	0.0810	1.0844	-0.0589	0.9427	0.4422	1.5561	0.2160	1.2411
hsa-mir-1	0.0630	1.0651	-0.1150	0.8913	-0.0637	0.9382	0.3702	1.3078	-0.0295	0.9708	-0.0208	0.9793	0.0089	1.0090	0.0954	1.1001	0.1580	1.1712	0.0084	1.0084
hsa-mir-30c	-0.0023	0.9976	-0.0550	0.9464	0.5053	1.6575	-0.0168	0.9833	0.0621	1.0641	-0.0121	0.9879	-0.4443	0.6412	0.1438	1.1547	-0.0056	0.9943	-0.6118	0.5423
hsa-mir-16	-0.0326	0.9678	0.0947	1.0993	-0.0634	0.9385	0.1409	1.1513	-0.0370	0.9636	0.0266	1.0270	0.0471	1.0482	-0.0732	0.9293	0.0058	1.0058	0.0447	1.0457
hsa-mir-30a	0.0845	1.0882	-0.1589	0.8530	0.8083	2.2441	0.0565	1.0582	0.0808	1.0841	-0.0779	0.9250	-0.1680	0.8452	0.0764	1.0794	0.0737	1.0765	-0.7061	0.4935
hsa-let-7i	0.1311	1.1401	0.4383	1.5501	0.1915	1.2110	0.2327	1.2621	-0.0354	0.9651	0.1394	1.1496	0.2574	1.2936	0.0702	1.0728	-0.1047	0.9005	0.0353	1.0360
hsa-mir-24	0.0191	1.0193	-0.0863	0.9172	0.0659	1.0681	0.4885	1.6298	0.0393	1.0401	-0.0345	0.9660	-0.0782	0.9246	0.0666	1.0688	0.0503	1.0516	-0.1175	0.8890
hsa-mir-95	-0.0345	0.9660	0.1185	1.1258	-0.0282	0.9721	0.2872	1.3326	0.0184	1.0186	0.1327	1.1419	0.0320	1.0325	0.0102	1.0102	0.0467	1.0478	-0.1247	0.8827

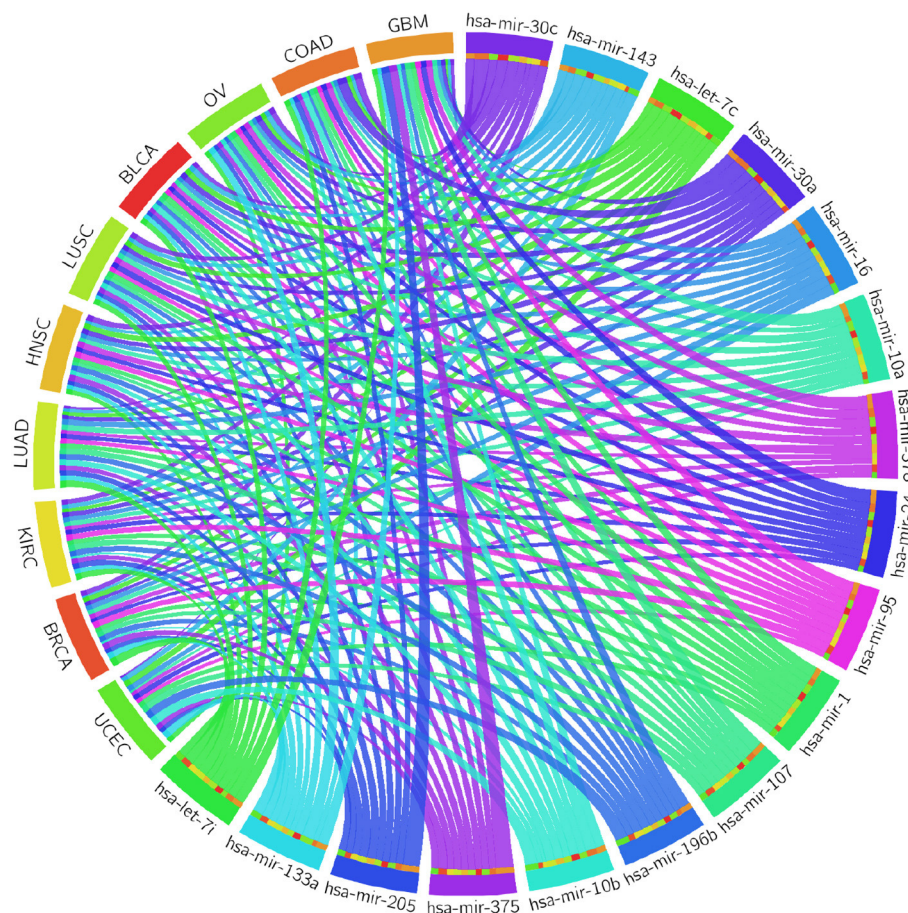


FIGURE 2 | Circos plot of Cox regression analysis results: Cox coefficient values are used to graphically visualize the association of 17 miRNAs with ten cancer types; a broader band signifies a stronger association between the miRNA and the particular cancer type.

hsa-mir-16, hsa-mir-30a, hsa-let-7i, hsa-mir-24, and hsa-mir-95 are highly associated with GBM, UCEC, OV, BLCA, COAD, and BRCA.

3.2.2. Network Analysis

For the 17 selected miRNAs, miRTarBase (Huang et al., 2020) was used to find their targets in order to elucidate their role in the different cancer types. To identify the most correlated targets, we computed the Pearson correlation between the expression values of miRNAs and mRNAs obtained from TCGA for the 10 cancer types and took the negative correlation value used in Zhou et al. (2015) as indicating strong association. The top five negatively correlated mRNAs associated with each of the 17 miRNAs are reported in **Table 6**; the rest are reported in **Supplementary Material**. To construct the interaction network, the miRNAs and their targets were ranked based on the cumulative negative correlation score and their presence in different cancer types as indicated by the association number. These results are reported in **Supplementary Table 3**. For example, hsa-mir-16 and its target, PHYHIP, are related to six cancer types and the cumulative negative correlation score is -4.142 . Similarly, hsa-mir-24 is correlated with C1QTNF6 in

another six cancer types, with a cumulative negative correlation score of -3.906 . The subset of such targeted mRNAs is used to construct the interaction network shown in **Figure 3**. The network reveals that the miRNAs hsa-mir-205, hsa-mir-10a, hsa-mir-107, hsa-mir-378, and hsa-mir-16 and their targets {CYR61, STARD8, TNFSF8}, {TTYH3, CARHSP1, LILRA2}, {CPEB3, TGFBR3, FGF2}, {NME4, NWD1, ORAI2}, and {PHYHIP, CPEB3, CTDSPL} play a crucial role in different types of cancer. The red, blue, pink, dark green, light green, and black edges in **Figure 3** signify that the number of cancer types associated with the corresponding pair of miRNA and target mRNA is 6, 5, 4, 3, 2, and 1, respectively. The targets of the miRNAs are investigated further using KEGG pathway, GO enrichment, and PPI network analysis in the following subsections in order to see their impact on the different types of cancer.

3.2.3. Expression Analysis

Expression analysis was conducted using a one-way ANOVA test in order to evaluate the statistical significance of the differential expression of the 17 selected miRNAs. Alternative techniques can also be used (Conesa et al., 2016; Costa-Silva et al., 2017; Crow et al., 2019). To perform the test, the population of patients

TABLE 5 | Cancer type most strongly associated with each selected miRNA, based on Cox coefficient.

miRNA	Cox coefficient	Hazard ratio	Cancer type	p-value	FDR	Regulation (up ↑ /down ↓)	PubMed ID
hsa-mir-205	0.9233	2.5177	GBM	5.14E-03	5.46E-03	↓	23054677
hsa-mir-10a	0.4881	1.6293	GBM	6.17E-24	1.87E-23	↓	20444541
hsa-mir-196b	0.4581	1.5811	UCEC	2.84E-48	1.21E-47	↑	–
hsa-mir-10b	0.5080	1.6621	UCEC	4.71E-03	4.71E-03	↓	–
hsa-mir-375	0.9882	1.7879	GBM	1.80E-14	2.36E-14	↓	29110584
hsa-mir-143	0.1504	1.1623	OV	3.38E-57	1.79E-56	↓	25304686
hsa-let-7c	0.1731	1.1890	BLCA	3.85E-38	1.31E-37	↓	21464941
hsa-mir-107	0.5363	1.7096	GBM	5.42E-24	1.87E-23	↑	24213470
hsa-mir-378	0.5511	1.7352	GBM	5.04E-20	7.79E-20	↓	29088758
hsa-mir-133a	0.4422	1.5561	OV	2.78E-20	4.73E-20	↑	24944666
hsa-mir-1	0.3702	1.3078	GBM	9.18E-08	1.04E-07	↑	–
hsa-mir-30c	0.5053	1.6575	COAD	4.48E-17	6.34E-17	↓	–
hsa-mir-16	0.1409	1.1513	GBM	5.42E-24	1.87E-23	↑	25864039
hsa-mir-30a	0.8083	2.2441	COAD	2.88E-56	9.80E-56	↓	22287560
hsa-let-7i	0.4383	1.5501	BRCA	1.33E-43	2.83E-43	↑	26378051
hsa-mir-24	0.4885	1.6298	GBM	5.42E-24	1.87E-23	↓	25864039
hsa-mir-95	0.2872	1.3326	GBM	6.59E-24	1.87E-23	↑	28155650

was divided into tumor and normal groups for a given miRNA in a particular cancer type. As a result of ANOVA, significant ($p < 0.05$) changes in expression were observed between the tumor and normal groups for the 17 selected miRNAs. For example, the p -values of hsa-mir-10a, hsa-mir-196b, hsa-mir-10b, hsa-mir-375, hsa-mir-143, hsa-let-7c, hsa-mir-107, hsa-mir-378, hsa-mir-133a, and hsa-mir-30c were 6.17E-24, 2.84E-48, 4.71E-03, 1.80E-14, 3.38E-57, 3.85E-38, 5.42E-24, 5.04E-20, 2.78E-20, and 4.48E-17 respectively. Additionally, box plots of the selected miRNAs in tumor and normal samples are provided in **Supplementary Figure 3**. To investigate the relationship between the expression levels of miRNAs for each cancer type, hierarchical clustering was performed on the tumor and normal samples of the 17 selected miRNAs. The results are shown in **Figure 4**, from which the change in expression levels of the miRNAs between tumor and normal samples is evident. In the cluster plots, red indicates high expression levels and green low expression levels; black corresponds to not significantly expressed samples.

3.2.4. KEGG Pathway Analysis

In order to perform KEGG pathway analysis for the 17 selected miRNAs in 10 cancer types, their targets were identified based on negative correlation values as described in section 3.2.2. Then, these target mRNAs were used in the DIANA tool (Vlachos et al., 2015) separately to identify significant KEGG pathways associated with the selected miRNAs in different cancer types. The five most significant pathways for each of the 17 miRNAs according to the FDR-corrected p -value within 5% statistical significance for the different cancer types are reported in **Table 7**. The detailed pathways of the 17 miRNAs with all their targets are presented in **Supplementary Material**. It can be seen from **Table 7** that the most significantly enriched pathways are

involved in various cancer types. For example, hsa-mir-205 and hsa-mir-133a are found to be enriched in pathways relating to *hsa05206: MicroRNAs in cancer* for nine cancer types. Similarly, hsa-mir-196b is found to be enriched in the pathway of *hsa05210: Colorectal cancer* for all 10 cancer types, with the FDR-corrected p -values of BLCA, BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, and UCEC being 2.24E-07, 2.30E-06, 2.31E-07, 7.00E-04, 2.36E-07, 4.54E-06, 8.48E-05, 2.38E-06, 2.24E-07, and 2.20E-07, respectively. In addition, critical pathways such as the *PI3K-Akt signaling pathway*, *p53 signaling pathway*, *Bladder cancer*, *Pancreatic cancer*, *Prostate cancer*, and *Lung cancer* are also found for the 17 miRNAs with FDR-corrected p -values within 5% statistical significance in 10 different cancer types. The presence of such critical pathways for the 17 selected miRNAs suggests that these miRNAs play a significant role in various cancer types, including the 10 types considered in this paper.

3.2.5. Gene Ontology Enrichment Analysis

Similar to the KEGG pathway analysis, GO enrichment analysis was also performed with the targets of the selected miRNAs using the Enrichr tool (Kuleshov et al., 2016), to assess the significance of the roles the selected miRNAs play in different biological activities. The results of different analyses for biological processes, molecular functions, and cellular components are reported in **Table 8** and in **Supplementary Tables 4, 5**, respectively; the details of the enrichment analysis are also given in **Supplementary Material**. In **Table 8**, we see that various biological process GO terms which are related to the targets of the 17 selected miRNAs have important roles in cancer development. For example, GO:0023051 is linked with regulation of signaling, which is involved in the development of colorectal cancer. Similarly, GO:0051252 is linked with regulation of the RNA metabolic process of bladder urothelial carcinoma,

TABLE 6 | Association of the 17 selected miRNAs and their top five targets in 10 cancer types.

miRNA	BLCA		BRCA		COAD		GBM		HNSC		KIRC		LUAD		LUSC		OV		UCEC	
	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score
hsa-mir-205	ZEB2	-7.45	E2F1	-3.29	SLC7A2	-9.95	PDLIM5	-5.28	RCAN2	-4.89	MAF	-4.68	ANGPTL7	-4.15	TRPV2	-6.43	PARD6B	-2.54	SLC7A2	-5.22
	ZEB1	-7.05	SATB2	-2.94	TIMP1	-9.63	ZNF707	-4.61	STARD8	-4.61	VEGFA	-4.58	ITM2A	-3.88	ALPK3	-6.20	PLCXD2	-2.10	SRC	-3.77
	SYT11	-7.02	RAB11FIP3	-2.88	SESN3	-9.43	ANKRD50	-4.25	ESRRG	-4.60	SLC37A4	-3.83	CPEB3	-3.54	STARD8	-6.16	C11orf74	-2.08	PARD6B	-3.55
	RCAN2	-6.99	ZFH3	-2.70	SAMD8	-8.99	ERBB3	-4.25	ZEB1	-4.55	FLCN	-3.79	FAM19A1	-3.47	ENPP4	-6.04	FGF2	-2.05	PISD	-3.23
	LRRK2	-6.99	CDK1	-2.63	HOXA11	-8.78	YES1	-4.14	CTGF	-4.40	FGFR1OP	-3.55	CTGF	-3.17	LPCAT1	-6.02	SLC39A14	-1.97	BCL9L	-3.13
hsa-mir-10a	NACC2	-5.33	E2F1	-3.93	H3F3C	-9.62	SLC2A3	-5.45	PANX1	-3.16	TTYH3	-6.29	TAF1D	-3.10	YOD1	-4.23	ARSK	-2.11	IRGQ	-5.32
	CLIC4	-5.29	BIRC5	-3.76	FHL2	-9.11	KIAA1143	-5.41	CARHSP1	-3.02	SCD	-6.28	DVL1	-2.93	ANP32E	-3.46	RIOK2	-2.04	FEM1A	-4.80
	COL6A2	-5.04	PPM1G	-3.69	MTR	-8.99	YOD1	-5.36	FHL2	-2.79	CD3D	-6.19	AHCYL2	-2.85	CHMP1B	-3.37	ZBTB10	-2.03	E2F1	-4.78
	RAP1A	-4.79	TIMM50	-3.60	SFT2D2	-8.82	BCL6	-5.18	TFAP2A	-2.67	KLHL6	-6.12	PABPC1	-2.77	HNRNPF	-3.20	CHL1	-2.01	NF2	-4.77
	TGFB3	-4.77	TPI1	-3.54	NF2	-8.79	DUSP3	-5.12	EBNA1BP2	-2.44	COL6A2	-6.03	YAP1	-2.65	NOP16	-3.13	TRA2B	-1.99	CHRNA5	-4.71
hsa-mir-196b	GATA6	-5.61	REEP5	-2.67	KCTD21	-9.89	MAP2K2	-5.92	PRUNE2	-3.57	HSD17B10	-2.28	MEIS1	-3.96	TGFB2	-7.15	MYC	-2.45	HOXB7	-3.46
	PRUNE2	-5.26	DCTN4	-2.18	ACER2	-9.74	MYC	-4.55	BEST3	-3.05	CALM1	-2.16	TBRG1	-3.85	LAMB2	-5.98	MARS2	-1.97	HOXB8	-3.17
	TGFB2	-4.74	PBX1	-2.03	BCAR3	-9.68	PRKACA	-4.42	IGDCC4	-2.82	PBX1	-1.86	REEP5	-3.80	TRPC3	-5.67	GATA6	-1.91	SLC23A2	-2.95
	NR4A3	-4.65	SUOX	-1.81	IGF2BP3	-9.43	IARS	-4.11	KLHDC8B	-2.59	C14orf37	-1.43	TRPC3	-3.53	NR4A3	-5.52	ALDOA	-1.78	TGFB3	-2.88
	SNX9	-4.59	TLE3	-1.78	HIST1H2BD	-9.10	HMGA1	-4.07	NR4A3	-2.33	SUOX	-1.41	TGFB2	-3.32	GATA6	-5.11	GGA3	-1.63	GATA6	-2.61
hsa-mir-10b	TPM1	-4.58	PLK1	-7.27	INHBA	-9.91	CMPK1	-6.23	RNF2	-2.50	TUBA1B	-5.41	MARVELD3	-3.34	LILRA2	-4.68	SDC1	-2.79	ASCL2	-3.78
	SFRP1	-4.13	BUB1	-6.83	MBNL3	-9.82	PDK3	-4.45	TTYH3	-2.43	HTATIP2	-5.11	NPEPPS	-2.92	AHCYL2	-4.52	INHBA	-2.72	GLB1L3	-3.53
	MBNL1	-4.08	CCNA2	-6.76	OPA3	-9.61	EXOSC2	-4.29	UBE2Z	-2.20	LILRB2	-4.95	TRIM2	-2.87	S1PR2	-4.17	CMPK1	-2.69	PLA2G2C	-2.95
	PPP3CB	-3.86	MELK	-6.76	SLC2A3	-9.55	HNRNPF	-4.28	SLC5A5	-2.15	LILRA2	-4.86	FAHD1	-2.59	PAG1	-3.95	TMED5	-2.26	GPCPD1	-2.94
	SGCD	-3.78	POC1A	-6.70	PPP1R13B	-9.35	EIF1	-4.21	FZD2	-2.09	PLK1	-4.86	ALKBH4	-2.57	FGD4	-3.74	SLC2A3	-2.14	MSTO1	-2.73
hsa-mir-375	CTGF	-4.17	FAM89A	-6.15	SON	-9.94	REEP3	-5.94	CALU	-4.71	HEY1	-3.62	JAG1	-5.77	PIK3CA	-4.28	DPYSL3	-2.98	KLHDC8B	-4.97
	FSTL3	-4.09	RHOQ	-5.97	LIMD2	-9.81	BAK1	-5.02	COL12A1	-4.50	ZNF785	-3.38	KLF4	-5.58	NUP54	-3.67	CLDN1	-2.50	ASAP2	-4.73
	TNS1	-4.09	CFL2	-5.87	ATG7	-9.80	SPRED1	-4.97	EXT1	-4.28	CDCA7L	-3.28	MBD2	-5.43	ARNTL2	-3.66	IL1RAP	-2.06	SFT2D2	-4.48
	SH3D19	-4.01	ACSL4	-5.86	JAK2	-9.58	CARD8	-4.64	CCDC88A	-4.12	CARD8	-3.15	AKAP7	-5.21	JAG1	-3.58	COL12A1	-2.03	CLDN1	-4.22
	SAMD4A	-3.99	CELF2	-5.71	ESPNL	-9.54	ZNF799	-4.58	NETO2	-4.04	NETO2	-3.13	CRIM1	-5.14	USP46	-3.46	SEC23A	-1.98	TSC22D2	-4.20
hsa-mir-143	OTUB1	-5.34	CENPM	-5.21	NKPD1	-9.87	HIST1H2BG	-4.67	SDC1	-2.79	FSD2	-2.98	TIMM8A	-5.13	COX6B1	-4.67	MAT2A	-2.64	FHIT	-3.62
	CAPZA1	-4.88	PIK3R2	-4.73	TIAL1	-9.81	ANG	-4.30	RAB22A	-2.43	DTNB	-2.81	RAB10	-5.01	TRUB2	-4.34	SLC25A33	-2.35	SNX22	-2.89
	SYNPO2L	-4.88	STXBP2	-4.69	TUBD1	-9.51	RER1	-3.97	TMEM40	-2.39	TMEM120B	-2.48	PRMT3	-5.00	RPS19	-4.06	RACGAP1	-2.26	C4orf19	-2.85
	STXBP2	-4.79	LMNB2	-4.68	PHAX	-9.48	GLB1L	-3.90	RAB10	-2.38	MRPS25	-2.47	PTCD3	-4.93	AKT2	-4.01	GPSM2	-2.22	RDH10	-2.39
	KCNA7	-4.65	AP1S1	-4.66	TTC38	-9.44	ADCY2	-3.88	NKPD1	-2.30	QPRT	-2.40	C15orf48	-4.89	OTUB1	-3.94	CAPZA1	-2.17	ZNF117	-2.32
hsa-let-7c	YWHAZ	-4.49	CCNF	-5.94	HMGXB4	-9.81	TRIB1	-6.35	ITGA3	-5.38	RRM2	-4.23	LDHA	-7.24	COX6B1	-3.43	CASP3	-2.68	TNFRSF10B	-3.26
	SLC20A1	-4.16	CKS2	-5.89	HES5	-9.80	ACTB	-6.14	LDHA	-5.05	DLX4	-4.04	EZH2	-6.88	NAA20	-3.32	E2F6	-2.44	SOD2	-3.18
	EFHD2	-3.93	CCNB2	-5.84	YWHAZ	-9.79	THBS1	-6.11	MT2A	-4.81	TNFSF9	-3.96	HMGA1	-6.79	SF3B4	-3.17	COIL	-2.35	RNF7	-2.95
	MRPL12	-3.86	RRM2	-5.72	WDR3	-9.67	SLC20A1	-5.80	HMGA2	-4.77	CCNB2	-3.96	EIF4A3	-6.74	KIAA0391	-3.14	RNFT1	-2.27	CEBPB	-2.64
	PCGF3	-3.73	H2AFZ	-5.65	LYN	-9.47	FHL2	-5.54	PSMB2	-4.28	FANCI	-3.79	MRPL12	-6.67	YWHAZ	-3.11	GABPB1	-2.24	ICOSLG	-2.49

(Continued)

TABLE 6 | Continued

miRNA	BLCA		BRCA		COAD		GBM		HNSC		KIRC		LUAD		LUSC		OV		UCEC	
	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score
hsa-mir-107	GLP2R	-7.83	CAV1	-8.34	SMARCA5	-9.64	REL	-4.65	CPEB3	-5.23	FOXC1	-8.31	RS1	-8.88	RS1	-8.75	SUN2	-4.14	CPEB1	-5.76
	PER1	-7.67	FGF2	-8.05	TMEM87A	-9.63	PIK3R1	-3.59	CHRM1	-4.70	PLAG1	-8.24	SH3GL2	-8.86	ALDH3B1	-6.89	ERN1	-2.97	CYSLTR2	-3.71
	CPEB1	-7.62	FOXO1	-7.82	CDC42SE2	-9.52	ZBTB38	-3.43	AMOT	-4.09	CPEB3	-8.13	TGFBR3	-8.45	LATS2	-6.77	VCAN	-2.96	FGF2	-3.41
	NFIA	-7.32	DST	-7.71	PURA	-9.42	CAV1	-3.36	TGFBR3	-3.51	SH3GL2	-8.09	CAV1	-8.26	PRKCE	-6.36	PAG1	-2.85	SLC28A1	-2.91
	DMPK	-7.00	KLF4	-7.62	UBE2Q1	-9.04	ADORA3	-3.08	NFIA	-3.36	CKMT1A	-7.93	FGF2	-8.15	PAG1	-6.34	YWHAH	-2.61	DAPK1	-2.73
hsa-mir-378	ENO1	-5.57	PTBP1	-6.96	TMEM154	-9.76	NWD1	-6.56	SERPINH1	-6.16	PRKD2	-5.07	BYSL	-6.95	MELK	-4.10	KLHL7	-3.51	MYOZ3	-3.84
	TTC4	-5.48	RABEP2	-6.75	MYADM	-9.69	RTN3	-6.27	ENAH	-5.83	PML	-4.96	ALDOA	-6.91	PRMT1	-3.81	IGDCC3	-3.48	IGSF3	-3.45
	MRPL37	-5.37	BBC3	-6.68	OPA3	-9.57	CYP2U1	-5.99	MARVELD1	-5.48	LDHA	-4.71	LDHA	-6.78	PTOV1	-3.50	ENAH	-3.25	NLGN2	-3.35
	CDK4	-5.27	HIST1H2BD	-6.65	UGT8	-9.55	WDR5B	-5.63	FBLIM1	-5.46	ORA12	-4.67	P4HB	-6.71	ENO1	-3.27	IGSF3	-3.19	MSC	-3.27
	FEN1	-5.08	DCTPP1	-6.34	KCNN1	-9.54	ENPP4	-5.50	MYO1B	-5.39	STOML1	-4.50	PAICS	-6.70	DCTPP1	-3.19	VAMP4	-3.16	YPEL1	-3.26
hsa-mir-133a	IGF2BP1	-3.91	MEG3	-5.74	TMEM59	-9.63	NR3C1	-4.30	PIGR	-2.49	PRDM16	-7.24	FERMT2	-3.94	PRDM16	-3.79	VEGFA	-2.21	PIGR	-2.42
	DEK	-3.52	PIGR	-5.39	UGT2B10	-9.30	ANGEL2	-4.28	PIAS2	-2.40	CMTM4	-6.94	ANGPT4	-3.84	MLEC	-3.72	AFTPH	-1.90	BCL3	-2.28
	CDK5R1	-3.49	PER2	-5.30	CDC42	-9.08	FAM160B1	-3.53	RBMXL1	-2.28	KCNQ1	-6.81	ZEB1	-3.75	ANGPT4	-3.53	CIAO1	-1.79	MYPN	-2.19
	SUPT16H	-3.43	NGFR	-4.77	SEC61B	-9.03	MMP14	-3.21	CIAO1	-2.23	ERBB2	-6.77	GRID1	-3.45	ARHGAP31	-3.50	KRT7	-1.72	KCNQ1	-2.08
	TCTEX1D2	-3.37	NR3C1	-4.65	MYL12A	-8.03	UBA2	-3.20	AFTPH	-2.23	PNP	-6.67	ZFP28	-3.30	ZEB1	-3.41	NR2C2	-1.64	NFAM1	-1.83
hsa-mir-1	NCAPG	-6.57	PTBP1	-5.38	BMP7	-9.88	SLC8A1	-5.20	LHX4	-3.32	VEGFA	-5.09	RFC5	-5.15	EFTUD2	-4.76	IFI44	-2.33	CD63	-2.85
	PTBP1	-6.33	ATP13A1	-4.69	CAST	-9.73	C1orf27	-5.03	FANCI	-3.32	ANGPTL4	-4.62	KIF4A	-4.92	IQGAP3	-4.72	MYEF2	-1.98	PPIB	-2.81
	RCC2	-6.23	OCIAD2	-4.54	CEBPA	-9.72	SCYL3	-4.94	SFXN1	-3.29	NETO2	-4.58	MAD2L1	-4.88	DSG2	-3.90	TWF1	-1.91	MTHFS	-2.72
	UHRF1	-6.21	KIAA1522	-4.54	GOLGA7	-9.67	LZTFL1	-4.80	BRI3BP	-3.27	KAT2A	-4.53	MTHFD2	-4.83	EIF4G1	-3.80	FOLR1	-1.90	MACROD1	-2.44
	SFXN1	-6.17	RCC2	-4.51	OAT	-9.63	WDR11	-4.78	NCAPD3	-3.01	HPS4	-4.44	SPC24	-4.76	SFXN1	-3.72	YTHDF2	-1.86	CRELD2	-2.24
hsa-mir-30c	PAM	-4.93	NRBP1	-3.56	SOC3	-9.78	MARCKSL1	-5.82	SNAI2	-3.65	VIM	-7.03	TOMM5	-4.76	BIRC5	-4.04	ETS1	-3.27	PSMD7	-4.26
	CXCL11	-4.92	RUNX2	-3.11	ARF3	-9.49	VASH1	-5.63	SERPINE1	-3.30	LHFPL2	-6.51	RPS3	-4.70	MYBL2	-4.03	MTDH	-3.15	UBE2I	-4.23
	SNAI2	-4.22	PRDM1	-2.94	FOXA1	-9.48	ETS1	-5.40	SLC7A5	-3.22	SH3GL1	-6.40	MRPS16	-4.43	SLC7A5	-3.91	ITGB3	-3.15	NDUFA12	-3.64
	ADAM9	-3.84	CASP3	-2.92	UBE2I	-9.40	DLL4	-5.37	CTSC	-3.04	IFNAR2	-6.33	BIRC5	-4.33	ECT2	-3.85	RRM2	-2.99	LLPH	-3.63
	MGAT2	-3.79	PHTF2	-2.82	CADPS2	-8.43	SOX12	-4.96	SLC38A7	-3.03	BIRC5	-6.22	UXT	-4.29	RRM2	-3.81	CASP3	-2.99	PPP2R1B	-3.58
hsa-mir-16	PHYHIP	-8.38	DMD	-8.17	ZYX	-9.96	HGF	-5.72	CPEB3	-5.13	DMRT2	-9.17	RS1	-8.88	DLC1	-9.04	SPRYD3	-2.93	PTPRT	-6.30
	NEGR1	-8.11	GNAL	-8.17	CLIP2	-9.96	CCPG1	-5.68	TMEM100	-4.95	C1orf226	-8.94	TGFBR3	-8.70	RS1	-8.82	ZCCHC3	-2.37	SLC6A4	-6.24
	GLP2R	-8.01	PLSCR4	-8.15	CAMSAP1	-9.95	IRAK3	-5.49	PDCD4	-4.56	SLC9A2	-8.89	WNT3A	-8.65	NEGR1	-8.32	CCND1	-2.22	PHYHIP	-5.34
	GNAL	-7.85	RBMS3	-8.10	TLL1	-9.93	PDXK	-5.45	PDK4	-4.51	SPTBN2	-8.87	SLC6A4	-8.54	SLC6A4	-8.09	BTRC	-2.19	PLSCR4	-5.18
	DIXDC1	-7.73	CDC14B	-7.83	KATNAL1	-9.90	PHYHIP	-5.42	ZBTB16	-4.30	SLC6A4	-8.84	AGER	-8.53	KDR	-7.97	BAMBI	-2.11	KCND3	-4.74
hsa-mir-30a	BAZ1B	-4.25	CBX2	-4.83	NCEH1	-9.64	PIK3R2	-5.50	FSCN1	-5.26	SERPINE1	-6.55	SFXN1	-6.36	KIF11	-6.08	C8orf76	-2.18	PPP2R1B	-4.41
	LMNB1	-4.22	CCNE2	-4.32	SBF1	-9.61	PES1	-5.43	FBXO45	-5.07	TGFB1	-6.26	PHTF2	-6.34	KPNA2	-5.87	MTHFD2	-2.07	GRPEL2	-4.39
	MYBL2	-4.16	RRM2	-4.23	SFXN1	-9.31	SIX4	-5.41	YWHAZ	-4.77	RUNX2	-6.24	PAICS	-6.24	CDC20	-5.84	CASP3	-2.01	DCTN4	-4.16
	RRM2	-4.10	C8orf76	-4.19	SOX12	-9.21	THOC5	-5.34	SLC16A1	-4.62	ITGA5	-6.15	TUBB3	-6.16	RRM2	-5.70	QRFPR	-1.96	IDH1	-4.11
	PAICS	-3.98	MYBL2	-4.10	RTP4	-9.16	NR2F6	-5.16	TUBB3	-4.57	CARS	-6.08	NUPL2	-6.14	PAICS	-5.56	CBX3	-1.75	STMN1	-4.06
	SYNJ2BP	-4.59	CRY2	-3.32	SLC20A1	-9.29	ONECUT2	-3.74	SDR42E1	-1.81	ANKRD46	-7.49	QDPR	-4.07	USP47	-3.18	HMG2A	-3.31	MEIS3P1	-3.39
	MYOCD	-4.31	MTUS1	-2.79	MSI2	-9.25	RABL2A	-3.70	CD59	-1.67	PAFAH2	-7.28	AHR	-3.23	GGA3	-3.10	ZCCHC3	-3.09	PBX2	-3.23

(Continued)

TABLE 6 | Continued

miRNA	BLCA		BRCA		COAD		GBM		HNSC		KIRC		LUAD		LUSC		OV		UCEC	
	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score	mRNA	Corr. score
hsa-let-7i	ACTA1	-4.30	DUSP1	-2.64	TUBB2A	-9.14	ODV3	-3.69	PRSS22	-1.46	TGFBR3	-6.99	ASCL1	-3.22	PIK3C2A	-3.05	COPS8	-2.98	SEMA4C	-2.55
	SEMA4C	-4.08	DNAH9	-2.53	SURF4	-8.86	NCKIPSD	-3.61	GLO1	-1.40	DNAJC28	-6.96	DNAH9	-3.17	TSC22D2	-2.93	PBX2	-2.59	RNF144B	-2.53
	KCNB1	-3.96	DNAJC28	-2.31	ACOT9	-8.67	C5orf51	-3.55	GRPEL2	-1.17	NAT8L	-6.96	TGFBR3	-3.02	IGF2BP1	-2.91	HMGAI	-2.57	DDOST	-2.51
	MEN1	-6.99	PKMYT1	-7.61	TLL7	-9.87	FLCN	-5.28	TGFBI	-6.65	NETO2	-8.35	CCNB1	-7.43	AURKB	-7.58	KAZALD1	-2.84	PRSS8	-2.95
hsa-mir-24	CDK1	-6.61	TRIM11	-7.45	DCAF4	-9.79	CYP20A1	-5.18	C1QTNF6	-6.46	STX4	-8.13	ALDOA	-7.41	UBE2C	-7.57	PLAGL2	-2.66	AGPAT2	-2.87
	ADPGK	-6.48	CCNB1	-7.28	OCL4	-9.73	ADPGK	-4.28	MMP14	-6.16	C1QTNF6	-8.08	RRM2	-7.39	CCNB1	-7.50	BTBD3	-2.44	PKMYT1	-2.67
	UBE2C	-6.45	UBE2C	-7.27	SIT1	-9.69	SMYD4	-4.27	FSCN1	-5.97	EHD2	-8.06	IMP4	-7.28	CCNA2	-7.29	ZNF107	-2.30	TNIP2	-2.62
	TRIM11	-6.44	CDK1	-7.23	OLR1	-9.63	MDM4	-4.18	NETO2	-5.94	CDKN2A	-8.03	LDHA	-7.25	RRM2	-7.04	DNAJC10	-2.23	ATL3	-2.58
hsa-mir-95	DGKB	-4.54	ITM2B	-3.83	TBX18	-9.74	CREBL2	-4.81	SCD	-2.30	RDH11	-3.52	FEM1C	-4.23	ZNF699	-3.27	ZNF460	-1.45	ITPR1P2	-3.77
	ACTC1	-4.27	CREBL2	-3.52	MRAS	-8.52	MCTS1	-4.80	RDH11	-2.20	MOSCS2	-3.41	ZBTB43	-4.15	PER1	-2.97	FEM1C	-1.36	CRK	-3.19
	FOXP2	-3.80	INTU	-3.50	ACVR1	-8.39	B3GNT2	-4.09	USP8	-2.07	METAP2	-3.16	ITM2B	-3.86	CDKN1A	-2.94	B3GNT2	-1.29	FAM126B	-3.06
	PTBP2	-3.79	ACVR1	-3.49	WARS	-8.15	TRIP4	-3.81	HFE	-1.98	ZNF711	-3.10	CREBL2	-3.84	REL	-2.87	DGKB	-1.26	FOXJ3	-3.01
	NXPH3	-3.65	SNX1	-3.35	ARPC1B	-8.10	CEBPD	-3.80	B3GNT2	-1.72	AVP11	-3.05	CELF2	-3.79	DUSP18	-2.80	CAONG8	-1.26	LRR058	-2.96

with an FDR-corrected p -value of 1.30E-03, which is <0.05 . Other GO terms for biological processes, such as GO:0051171, GO:0048519, and GO:0009653, are linked with the regulation of nitrogen compound metabolic process, negative regulation of biological process, and anatomical structure morphogenesis, respectively, for different cancer types.

As with the biological processes, GO terms for molecular functions were also found to be significant in the development of various types of cancer, as reported in **Supplementary Table 4**. For example, GO:0008134 is linked with transcription factor binding, which plays an important role in BRCA, GBM, KIRC, LUAD, OV, and UCEC, with respective FDR-corrected p -values 8.10E-03, 8.31E-06, 6.80E-06, 4.70E-03, 3.76E-05, and 3.40E-03 all being <0.05 . Other important GO terms such as GO:0044877, GO:0003723, GO:0003676, and GO:0008092 are found to be linked to protein-containing complex binding, RNA binding, nucleic acid binding, and cytoskeletal protein binding, respectively, for different cancer types. **Supplementary Table 5** reports the significant GO terms for cellular components in various cancer types. GO:0070013 is linked to intracellular organelle lumen, which has FDR-corrected p -values within the 5% significance level for seven different cancer types, namely BLCA, BRCA, COAD, GBM, HNSC, KIRC, and LUSC. In addition, GO:0043227, GO:0005654, and GO:0044444 are associated with membrane-bounded organelle, nucleoplasm, and cytoplasmic part, respectively, which are critical processes in the progression of different types of cancer. The relationship between these GO terms and important activities in cancer development have also been cross-validated in other studies (Waldman et al., 1997; Dhillon et al., 2007; He et al., 2013; Reimand et al., 2013; McClurg and Robson, 2015). Taken together, all of these evidences point to the potential importance of the 17 selected miRNAs in the development of various types of cancer.

3.2.6. Protein-Protein Interaction Network Analysis

In PPI networks, a node and an edge signify the interaction of a given protein and the protein-protein association. Here, related proteins share common functions, although they do not necessarily physically interact with each other. In our study, to perform the PPI network analysis, 170 sets of targets relating to 17 miRNAs in 10 different cancer types were used to compute the PPI networks using the STRING database (Szklarczyk et al., 2019). Then, the 170 interaction networks were further analyzed to rank the proteins based on the degree of their nodes and their presence in 10 cancer types. The top 30 proteins are reported in **Table 9**, while the rest are given in **Supplementary Material**. For example, MYC has degrees 34, 28, 41, 85, 33, 133, 54, 25, 38, and 32 with respect to BLCA, BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, and UCEC, respectively. Therefore, the total degree of MYC is 503. Similarly, other proteins have certain degrees in different cancer types. Moreover, their presence in the different cancer types is indicated by the association count; for example, MYC has an association count of 10. The proteins in **Table 9** are used to construct the final consolidated PPI network shown in **Figure 5**, which represents the associations between the top 30 proteins and 10 different cancer types. The average node

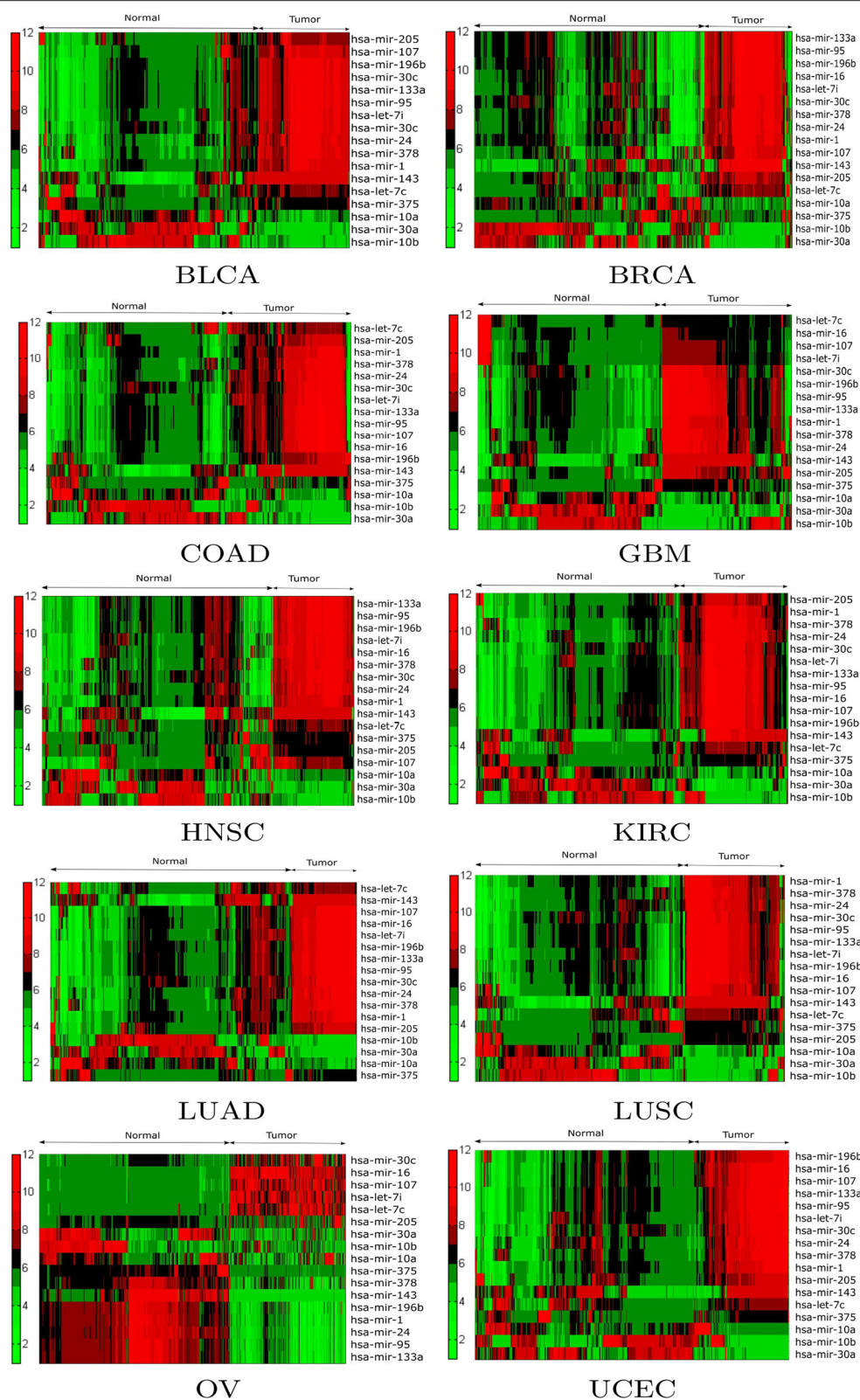


FIGURE 4 | Hierarchical clustering results of the differentially expressed miRNAs for the BLCA, BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, and UCEC datasets. Red indicates high expression levels, green low expression levels, and black not significantly expressed samples.

TABLE 7 | Five significant KEGG pathways for each of the 17 selected miRNAs in 10 cancer types.

miRNA	Pathway	FDR-corrected <i>p</i> -value									
		BLCA	BRCA	COAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC
hsa-mir-205	hsa05206: MicroRNAs in cancer	2.20E-03	–	1.60E-04	1.73E-05	2.23E-02	1.02E-06	2.20E-04	1.70E-04	2.00E-03	2.40E-03
	hsa05202: Transcriptional misregulation in cancer	1.33E-02	–	1.62E-02	–	–	1.11E-02	2.15E-02	–	9.20E-03	–
	hsa05219: Bladder cancer	–	1.06E-02	–	–	–	3.54E-02	–	–	–	6.30E-03
	hsa05205: Proteoglycans in cancer	–	–	–	2.38E-02	–	1.73E-02	–	–	–	6.30E-03
	hsa04520: Adherens junction	2.13E-02	–	–	–	–	–	2.80E-02	–	–	–
hsa-mir-10a	hsa05016: Huntington's disease	1.27E-02	–	–	–	–	–	–	4.64E-02	–	–
	hsa04218: Cellular senescence	–	–	–	–	2.90E-03	3.01E-02	–	–	–	–
	hsa04714: Thermogenesis	1.27E-02	–	–	–	–	–	–	–	–	–
	hsa04510: Focal adhesion	1.27E-02	–	–	–	–	–	–	–	–	–
	hsa04213: Longevity regulating pathway—multiple species	1.27E-02	–	–	–	–	–	–	–	–	–
hsa-mir-196b	hsa05210: Colorectal cancer	2.24E-07	2.30E-06	2.31E-07	7.00E-04	2.36E-07	4.54E-06	8.48E-05	2.38E-06	2.24E-07	2.20E-07
	hsa01522: Endocrine resistance	2.77E-07	5.58E-07	–	–	2.92E-07	5.08E-06	8.58E-05	5.75E-07	–	–
	hsa05215: Prostate cancer	2.39E-06	3.00E-06	–	–	2.51E-06	3.81E-05	–	3.09E-06	3.42E-06	–
	hsa05161: Hepatitis B	–	3.00E-06	–	1.52E-05	3.61E-07	–	8.58E-05	3.09E-06	5.15E-07	–
	hsa04915: Estrogen signaling pathway	2.15E-06	2.88E-06	–	–	2.26E-06	3.27E-05	–	2.97E-06	–	–
hsa-mir-10b	hsa05169: Epstein-Barr virus infection	–	3.28E-02	–	2.40E-03	2.90E-02	1.10E-04	–	–	1.19E-02	–
	hsa04550: Signaling pathways regulating pluripotency of stem cells	–	–	1.50E-03	–	1.48E-02	–	–	–	1.19E-02	–
	hsa05206: MicroRNAs in cancer	2.50E-03	–	–	–	–	–	–	–	1.19E-02	–
	hsa04110: Cell cycle	–	9.10E-03	–	–	–	2.80E-04	–	–	–	–
	hsa04914: Progesterone-mediated oocyte maturation	–	1.34E-02	–	–	4.95E-02	–	–	–	–	–
hsa-mir-375	hsa01521: EGFR tyrosine kinase inhibitor resistance	7.00E-04	1.09E-02	–	–	–	–	–	–	7.50E-04	2.80E-03
	hsa04550: Signaling pathways regulating pluripotency of stem cells	4.02E-02	1.09E-02	–	–	–	–	–	–	–	–
	hsa04066: HIF-1 signaling pathway	–	–	2.32E-02	–	–	–	–	4.60E-03	–	–
	hsa05165: Human papillomavirus infection	–	–	2.32E-02	8.70E-03	–	–	–	–	–	–
	hsa05224: Breast cancer	–	–	–	6.00E-03	–	1.90E-03	–	–	–	–
hsa-mir-143	hsa05230: Central carbon metabolism in cancer	–	1.00E-03	–	–	4.09E-02	–	–	–	–	–
	hsa05213: Endometrial cancer	–	1.00E-03	–	–	–	–	–	–	–	–
	hsa05206: MicroRNAs in cancer	–	1.00E-03	–	–	–	–	–	–	–	–
	hsa05205: Proteoglycans in cancer	–	1.00E-03	–	–	–	–	–	–	–	–
	hsa05161: Hepatitis B	–	1.00E-03	–	–	–	–	–	–	–	–

(Continued)

TABLE 7 | Continued

miRNA	Pathway	FDR-corrected <i>p</i> -value									
		BLCA	BRCA	COAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC
hsa-let-7c	hsa05206: MicroRNAs in cancer	–	4.05E-02	1.85E-02	2.20E-04	–	1.10E-04	–	–	–	–
	hsa04115: p53 signaling pathway	–	–	–	2.90E-03	–	1.63E-05	4.43E-02	–	–	–
	hsa04110: Cell cycle	–	6.10E-03	–	–	–	–	4.43E-02	–	–	–
	hsa05222: Small cell lung cancer	–	4.05E-02	–	–	–	4.10E-04	–	–	–	–
	hsa04215: Apoptosis—multiple species	–	4.05E-02	–	–	–	–	4.43E-02	–	–	–
hsa-mir-107	hsa05200: Pathways in cancer	2.50E-03	1.30E-03	–	1.31E-02	–	–	2.40E-03	4.70E-04	7.38E-05	–
	hsa01521: EGFR tyrosine kinase inhibitor resistance	5.50E-03	1.30E-03	8.20E-03	–	–	–	5.20E-03	5.60E-03	3.50E-04	–
	hsa05165: Human papillomavirus infection	8.60E-03	1.30E-03	1.41E-02	–	–	–	–	8.80E-03	2.30E-03	–
	hsa04151: PI3K-Akt signaling pathway	5.50E-03	1.40E-03	–	–	–	–	–	5.60E-03	2.70E-03	–
	hsa05224: Breast cancer	5.50E-03	–	–	–	–	–	5.30E-03	–	–	–
hsa-mir-378	hsa03013: RNA transport	1.49E-02	–	–	–	–	–	3.11E-02	2.04E-02	–	–
	hsa03010: Ribosome	–	–	–	–	–	–	6.60E-04	2.04E-02	–	–
	hsa01100: Metabolic pathways	–	–	–	–	–	–	2.56E-02	2.04E-02	–	–
	hsa00010: Glycolysis/Gluconeogenesis	–	–	–	–	–	–	2.56E-02	2.04E-02	–	–
	hsa05224: Breast cancer	–	–	–	–	1.25E-02	–	–	–	–	–
hsa-mir-133a	hsa05206: MicroRNAs in cancer	7.80E-03	5.80E-03	1.76E-05	1.60E-04	4.70E-03	2.21E-02	1.92E-05	1.40E-06	7.20E-03	–
	hsa05215: Prostate cancer	7.80E-03	1.60E-04	–	3.90E-03	7.90E-04	2.21E-02	–	8.00E-04	7.20E-03	–
	hsa05212: Pancreatic cancer	5.50E-04	–	4.20E-03	–	5.50E-04	2.21E-02	–	–	6.20E-03	–
	hsa01524: Platinum drug resistance	3.30E-03	–	2.71E-02	–	1.70E-03	–	–	–	6.20E-03	–
	hsa05205: Proteoglycans in cancer	–	–	–	3.10E-03	–	–	8.34E-05	8.00E-04	–	–
hsa-mir-1	hsa03030: DNA replication	1.34E-10	1.83E-07	–	–	6.25E-06	7.60E-03	6.10E-09	1.16E-05	–	–
	hsa03430: Mismatch repair	3.80E-04	7.10E-03	–	–	3.10E-04	–	4.00E-04	1.75E-02	–	–
	hsa04110: Cell cycle	2.50E-04	9.20E-03	–	–	2.10E-04	–	8.74E-08	–	–	–
	hsa03015: mRNA surveillance pathway	3.64E-02	–	–	–	–	–	–	1.01E-02	–	–
	hsa05166: HTLV-I infection	4.42E-02	–	–	–	–	–	–	–	–	–
hsa-mir-30c	hsa05206: MicroRNAs in cancer	1.79E-05	–	2.90E-03	–	–	2.04E-05	–	–	1.21E-02	–
	hsa05200: Pathways in cancer	–	8.81E-05	–	2.40E-03	2.07E-02	3.40E-03	–	–	–	–
	hsa05211: Renal cell carcinoma	–	2.30E-03	–	2.09E-02	6.40E-03	–	1.42E-02	–	–	–
	hsa04141: Protein processing in endoplasmic reticulum	–	2.30E-03	–	–	1.93E-02	–	1.04E-02	2.56E-02	–	–
	hsa04380: Osteoclast differentiation	2.71E-02	–	5.10E-03	–	–	–	–	–	9.20E-03	–
hsa-mir-16	hsa04510: Focal adhesion	1.90E-03	5.80E-03	–	–	–	–	–	–	–	–
	hsa04010: MAPK signaling pathway	1.31E-02	6.30E-03	–	–	–	–	–	–	–	–
	hsa05200: Pathways in cancer	1.58E-02	–	–	–	–	–	7.42E-05	–	–	–
	hsa04151: PI3K-Akt signaling pathway	–	2.60E-03	–	–	–	–	–	–	3.20E-03	–
	hsa05206: MicroRNAs in cancer	–	6.30E-03	–	–	–	–	7.40E-04	–	–	–

(Continued)

TABLE 7 | Continued

miRNA	Pathway	FDR-corrected <i>p</i> -value									
		BLCA	BRCA	COAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC
hsa-mir-30a	hsa04110: Cell cycle	–	4.39E-06	–	3.90E-03	–	5.81E-05	2.30E-03	1.93E-06	–	–
	hsa05206: MicroRNAs in cancer	–	–	–	2.10E-04	2.28E-02	6.20E-04	–	5.70E-03	–	–
	hsa05203: Viral carcinogenesis	–	5.50E-04	–	–	–	1.80E-03	–	–	2.14E-02	–
	hsa05130: Pathogenic <i>Escherichia coli</i> infection	–	1.24E-02	–	–	2.28E-02	–	–	5.70E-03	–	–
	hsa03013: RNA transport	–	–	–	–	–	–	2.00E-03	5.70E-03	–	–
hsa-let-7i	hsa05206: MicroRNAs in cancer	–	–	9.80E-03	–	1.97E-02	–	1.80E-04	–	–	–
	hsa04550: Signaling pathways regulating pluripotency of stem cells	1.20E-03	–	–	–	–	–	6.20E-04	–	–	–
	hsa04066: HIF-1 signaling pathway	1.20E-03	–	–	–	–	–	6.80E-04	–	–	–
	hsa05130: Pathogenic <i>Escherichia coli</i> infection	–	–	1.60E-03	–	1.97E-02	–	–	–	–	–
	hsa05225: Hepatocellular carcinoma	1.01E-02	–	–	–	–	–	–	–	–	–
hsa-mir-24	hsa04110: Cell cycle	1.29E-06	1.57E-07	–	–	9.85E-09	5.43E-05	4.19E-10	4.40E-10	–	–
	hsa03030: DNA replication	2.13E-07	3.59E-06	–	–	1.10E-04	–	1.20E-03	1.03E-07	–	–
	hsa04218: Cellular senescence	–	4.00E-03	–	–	1.10E-04	1.50E-04	9.90E-04	4.20E-04	–	–
	hsa04115: p53 signaling pathway	7.49E-05	1.30E-03	–	–	8.40E-04	–	1.20E-03	–	–	–
	hsa00240: Pyrimidine metabolism	4.50E-04	–	–	–	–	–	–	4.28E-05	–	–
hsa-mir-95	hsa05211: Renal cell carcinoma	4.20E-03	–	4.93E-02	–	–	–	4.30E-03	4.10E-03	–	–
	hsa05220: Chronic myeloid leukemia	–	–	4.93E-02	–	–	–	2.26E-02	–	–	–
	hsa05219: Bladder cancer	–	–	4.93E-02	–	–	–	4.90E-03	–	–	–
	hsa05206: MicroRNAs in cancer	–	–	4.93E-02	–	–	–	2.79E-02	–	–	–
	hsa05203: Viral carcinogenesis	–	–	4.93E-02	–	–	–	1.67E-02	–	–	–

TABLE 8 | Five significant GO biological processes for each of the 17 selected miRNAs in 10 cancer types.

miRNA	GO biological process	FDR-corrected p-value									
		BLCA	BRCA	COAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC
hsa-mir-205	GO:0051239 regulation of multicellular organismal process	1.30E-04	–	4.34E-05	–	–	–	–	4.70E-04	–	–
	GO:0023051 regulation of signaling	–	–	6.76E-05	–	6.10E-04	–	–	4.70E-04	–	–
	GO:0010646 regulation of cell communication	–	–	6.76E-05	–	7.30E-04	–	–	4.70E-04	–	–
	GO:0060255 regulation of macromolecule metabolic process	–	–	–	3.87E-07	–	7.21E-06	–	–	–	7.59E-05
	GO:0051171 regulation of nitrogen compound metabolic process	–	–	–	3.87E-07	–	2.61E-05	–	–	–	7.59E-05
hsa-mir-10a	GO:0006139 nucleobase-containing compound metabolic process	–	8.72E-05	–	–	–	–	2.90E-04	–	1.10E-04	–
	GO:0071840 cellular component organization or biogenesis	2.33E-02	–	–	–	1.39E-06	–	–	–	–	–
	GO:0016043 cellular component organization	2.33E-02	–	–	–	7.28E-05	–	–	–	–	–
	GO:0034641 cellular nitrogen compound metabolic process	–	8.72E-05	–	–	–	–	–	4.60E-03	–	–
	GO:0010467 gene expression	–	1.70E-04	–	–	–	–	–	8.00E-04	–	–
hsa-mir-196b	GO:0048523 negative regulation of cellular process	8.93E-06	–	1.92E-07	2.60E-04	–	6.77E-05	1.30E-04	2.62E-06	1.20E-04	–
	GO:0048519 negative regulation of biological process	2.58E-05	–	1.92E-07	2.60E-04	–	6.77E-05	–	5.64E-06	–	–
	GO:0070482 response to oxygen levels	–	–	–	–	5.06E-06	–	4.43E-05	1.18E-05	1.20E-04	–
	GO:1901700 response to oxygen-containing compound	–	–	–	–	1.00E-04	–	–	1.11E-05	3.28E-05	4.54E-05
	GO:0051173 positive regulation of nitrogen compound metabolic process	2.58E-05	–	1.62E-06	–	–	–	–	–	1.10E-04	–
hsa-mir-10b	GO:0009653 anatomical structure morphogenesis	2.70E-04	–	3.12E-05	–	–	–	–	1.20E-04	–	–
	GO:1901564 organonitrogen compound metabolic process	–	–	–	2.80E-04	9.30E-03	9.50E-03	–	–	–	–
	GO:0048468 cell development	2.50E-04	–	1.10E-04	–	–	–	–	–	–	–
	GO:1903047 mitotic cell cycle process	–	3.30E-03	–	–	2.65E-02	–	–	–	–	–
	GO:0044237 cellular metabolic process	–	–	–	7.10E-04	–	7.00E-03	–	–	–	–
hsa-mir-375	GO:0048522 positive regulation of cellular process	–	1.25E-05	5.30E-03	–	–	–	5.46E-07	–	5.60E-04	5.50E-03
	GO:0031325 positive regulation of cellular metabolic process	–	1.51E-05	7.60E-03	–	–	–	5.82E-06	–	–	–
	GO:1902533 positive regulation of intracellular signal transduction	5.80E-03	–	–	–	–	–	–	–	5.60E-04	–
	GO:0051173 positive regulation of nitrogen compound metabolic process	–	1.51E-05	–	–	3.61E-02	–	–	–	–	–
	GO:0071840 cellular component organization or biogenesis	–	–	8.40E-03	–	–	–	–	–	5.60E-04	–
hsa-mir-143	GO:0044237 cellular metabolic process	–	2.84E-02	–	–	–	–	–	–	2.30E-03	–
	GO:0008152 metabolic process	–	2.84E-02	–	–	–	–	–	–	8.40E-03	–
	GO:1905477 positive regulation of protein localization to membrane	–	2.84E-02	–	–	–	–	–	–	–	–
	GO:1903829 positive regulation of cellular protein localization	–	2.84E-02	–	–	–	–	–	–	–	–
	GO:0006807 nitrogen compound metabolic process	–	2.84E-02	–	–	–	–	–	–	–	–

(Continued)

TABLE 8 | Continued

miRNA	GO biological process	FDR-corrected <i>p</i> -value									
		BLCA	BRCA	COAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC
hsa-let-7c	GO:0071840 cellular component organization or biogenesis	–	–	4.50E-03	–	6.50E-04	–	–	7.68E-06	–	–
	GO:0016043 cellular component organization	–	–	5.70E-03	–	6.50E-04	–	–	1.49E-05	–	–
	GO:0010604 positive regulation of macromolecule metabolic process	–	–	–	3.95E-09	–	–	–	1.20E-04	1.50E-03	–
	GO:0090304 nucleic acid metabolic process	4.30E-04	–	–	–	–	–	–	4.41E-05	–	–
	GO:0051252 regulation of RNA metabolic process	1.30E-03	–	–	–	–	–	–	–	–	2.16E-02
hsa-mir-107	GO:0048522 positive regulation of cellular process	5.54E-07	5.57E-06	5.00E-03	8.70E-03	–	–	–	1.44E-07	1.40E-04	–
	GO:0048519 negative regulation of biological process	–	–	–	8.70E-03	9.40E-04	3.38E-02	2.88E-06	–	–	–
	GO:0051172 negative regulation of nitrogen compound metabolic process	5.54E-07	–	–	–	–	–	2.31E-06	–	6.70E-05	–
	GO:0048518 positive regulation of biological process	1.17E-06	1.07E-05	–	–	–	–	–	2.42E-07	–	–
	GO:0080090 regulation of primary metabolic process	1.31E-06	–	–	–	–	–	–	–	–	9.80E-04
hsa-mir-378	GO:0044237 cellular metabolic process	–	1.07E-02	–	–	–	–	4.46E-05	2.30E-04	–	–
	GO:1901576 organic substance biosynthetic process	–	1.07E-02	–	–	–	–	4.46E-05	–	–	–
	GO:0044249 cellular biosynthetic process	–	1.07E-02	–	–	–	–	4.46E-05	–	–	–
	GO:0034645 cellular macromolecule biosynthetic process	–	1.07E-02	–	–	–	–	–	–	–	1.25E-02
	GO:0009058 biosynthetic process	–	1.07E-02	–	–	–	–	1.45E-05	–	–	–
hsa-mir-133a	GO:0071495 cellular response to endogenous stimulus	–	4.66E-05	–	–	–	2.21E-05	7.70E-04	–	8.10E-03	–
	GO:1901701 cellular response to oxygen-containing compound	–	4.40E-04	–	–	–	3.10E-04	7.70E-04	–	–	–
	GO:0071417 cellular response to organonitrogen compound	–	7.30E-04	–	–	8.70E-03	3.10E-04	–	–	–	–
	GO:0048518 positive regulation of biological process	–	–	–	5.00E-03	8.70E-03	–	–	–	–	9.23E-05
	GO:0065008 regulation of biological quality	–	–	1.50E-03	–	–	–	7.70E-04	–	–	–
hsa-mir-1	GO:0007049 cell cycle	1.76E-09	8.75E-06	–	–	6.45E-08	–	2.12E-06	4.40E-04	–	9.80E-04
	GO:0051276 chromosome organization	1.52E-07	3.59E-05	–	–	9.12E-07	–	5.41E-07	7.92E-05	–	–
	GO:0006261 DNA-dependent DNA replication	4.12E-08	1.40E-04	–	–	–	–	6.25E-07	–	–	–
	GO:0022402 cell cycle process	4.03E-07	–	–	–	2.18E-06	–	–	1.70E-03	–	–
	GO:0006259 DNA metabolic process	–	1.70E-04	–	–	5.77E-06	–	5.41E-07	–	–	–
hsa-mir-30c	GO:0010033 response to organic substance	–	1.03E-02	5.58E-05	–	–	–	–	–	1.97E-05	–
	GO:0060255 regulation of macromolecule metabolic process	3.20E-04	–	–	1.20E-04	–	–	–	–	–	–
	GO:0070887 cellular response to chemical stimulus	–	1.03E-02	2.00E-03	–	–	–	–	–	–	–
	GO:0016192 vesicle-mediated transport	–	1.03E-02	–	–	–	3.60E-03	–	–	–	–
	GO:0002376 immune system process	–	1.03E-02	–	–	–	–	2.05E-05	–	–	–
hsa-mir-16	GO:0051239 regulation of multicellular organismal process	4.64E-05	5.02E-08	–	–	–	–	–	–	–	–
	GO:0048731 system development	4.10E-04	–	–	–	–	–	5.33E-08	–	–	–
	GO:0045595 regulation of cell differentiation	4.10E-04	–	–	–	–	–	2.61E-08	–	–	–
	GO:0050793 regulation of developmental process	–	1.11E-07	–	–	–	–	2.61E-08	–	–	–

(Continued)

TABLE 8 | Continued

miRNA	GO biological process	FDR-corrected <i>p</i> -value									
		BLCA	BRCA	COAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC
	GO:2000026 regulation of multicellular organismal development	–	1.29E-07	–	–	–	–	–	–	1.50E-03	–
	GO:0071840 cellular component organization or biogenesis	–	–	–	–	1.30E-04	–	6.20E-03	1.02E-06	–	2.04E-02
	GO:1903047 mitotic cell cycle process	–	1.99E-02	–	–	–	–	6.20E-03	8.40E-06	–	–
hsa-mir-30a	GO:0090304 nucleic acid metabolic process	–	–	–	6.20E-04	–	–	–	–	–	2.73E-02
	GO:0071310 cellular response to organic substance	–	–	–	–	–	1.90E-03	–	–	3.20E-03	–
	GO:0006260 DNA replication	–	–	–	–	–	–	–	6.33E-06	–	2.04E-02
	GO:0080090 regulation of primary metabolic process	–	–	–	–	–	–	1.35E-02	2.20E-03	–	–
	GO:2000727 positive regulation of cardiac muscle cell differentiation	1.05E-02	–	–	–	–	–	–	–	–	–
hsa-let-7i	GO:1904705 regulation of vascular smooth muscle cell proliferation	1.05E-02	–	–	–	–	–	–	–	–	–
	GO:0071900 regulation of protein serine/threonine kinase activity	1.05E-02	–	–	–	–	–	–	–	–	–
	GO:0061061 muscle structure development	1.05E-02	–	–	–	–	–	–	–	–	–
hsa-mir-24	GO:0007049 cell cycle	1.13E-13	1.08E-08	–	–	2.65E-11	–	1.14E-08	1.16E-13	–	2.32E-09
	GO:0000278 mitotic cell cycle	1.72E-13	1.08E-08	–	–	4.26E-13	7.08E-05	2.78E-10	1.17E-12	–	–
	GO:1903047 mitotic cell cycle process	1.16E-11	4.56E-08	–	–	1.04E-10	7.08E-05	6.94E-10	1.17E-12	–	–
	GO:0022402 cell cycle process	3.15E-11	–	–	–	5.20E-11	9.55E-05	2.75E-07	9.89E-13	–	–
	GO:0044772 mitotic cell cycle phase transition	2.15E-10	4.56E-08	–	–	–	–	4.86E-10	1.51E-11	–	–
hsa-mir-95	GO:0060255 regulation of macromolecule metabolic process	3.22E-06	2.03E-05	8.31E-07	6.89E-05	1.30E-03	5.67E-06	2.05E-05	1.88E-06	1.03E-05	9.40E-04
	GO:0080090 regulation of primary metabolic process	3.22E-06	2.03E-05	8.31E-07	–	–	–	5.80E-05	2.99E-06	1.19E-05	9.40E-04
	GO:0051171 regulation of nitrogen compound metabolic process	3.22E-06	2.03E-05	8.31E-07	–	–	9.97E-06	5.80E-05	2.99E-06	1.19E-05	–
	GO:0050789 regulation of biological process	9.66E-07	–	1.93E-07	6.89E-05	1.30E-03	9.97E-06	–	–	4.16E-06	–
	GO:0065007 biological regulation	3.22E-06	–	8.31E-07	6.89E-05	1.30E-03	–	–	–	1.19E-05	9.40E-04

TABLE 9 | Association of top 30 proteins in 10 cancer types for the 17 selected miRNAs through their targets.

TF	Node degree of protein in 10 cancer types										Total degree	Association count
	BLCA	BRCA	COAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC		
MYC	34	28	41	85	33	133	54	25	38	32	503	10
VEGFA	23	10	15	53	23	18	36	36	48	39	301	10
AKT1	17	59	50	16	54	32	0	16	17	8	269	9
RRM2	17	28	0	11	21	23	30	32	10	10	182	9
CDK1	23	31	10	13	21	0	21	30	0	24	173	8
CDKN1A	20	19	17	15	15	18	20	17	8	10	159	10
UHRF1	18	29	1	7	21	23	25	14	0	5	143	9
CHEK1	24	22	0	0	21	0	24	28	0	22	141	6
H2AFX	32	16	9	0	10	0	16	21	8	24	136	8
MCM10	20	22	0	0	19	11	20	23	0	18	133	7
POLD1	33	26	0	0	15	14	16	30	0	0	134	6
IL6	11	12	7	29	8	0	15	15	14	16	127	9
RHOA	11	15	9	0	0	8	30	18	22	12	125	8
PCNA	20	26	0	0	0	0	25	40	0	17	128	5
DTL	13	18	0	3	12	13	13	23	10	16	121	9
CCNF	6	18	5	14	7	18	18	14	19	0	119	9
BRCA1	27	12	0	0	25	12	7	25	12	0	120	7
CDC42	8	0	22	17	0	13	12	23	0	19	114	7
PTEN	15	10	7	9	3	0	25	25	18	0	112	8
YWHAZ	5	12	13	9	27	4	11	13	7	3	104	10
PAICS	7	19	0	5	9	0	23	27	1	4	95	8
PIK3R1	7	19	7	7	10	0	16	8	9	5	88	9
UBA52	11	0	8	10	0	16	18	16	0	8	87	7
CTGF	11	12	10	8	5	4	16	13	5	0	84	9
KIF4A	17	15	0	0	10	8	15	11	10	0	86	7
MTOR	9	8	9	0	9	14	7	9	8	8	81	9
UBE2C	13	13	0	0	11	8	12	15	0	11	83	7
KIF2C	16	15	0	0	11	0	16	12	9	0	79	6
KIF18B	11	11	0	0	11	8	11	12	0	11	75	7
CHAF1B	10	13	0	4	12	0	12	9	5	2	67	8

covariance matrix adaptation evolutionary strategy (CMA-ES), forward selection (FS), and a classification technique. Features are reordered using SNE by performing clustering of highly correlated features. From the result of the clustering, a subset of features is randomly selected to perform multi-class classification on 10 cancer types. Although the features are randomly selected, the underlying classification task is treated as an optimization problem for CMA-ES in order to find the features automatically. Thereafter, a final set of features/miRNAs is obtained using forward selection. The results of the first part of SCES-FS have been compared with the results of some well-known feature selection methods, including ESVM-RFE, LASSO, NSGA-II-SE, MOGA, SVM-nRFE, SVM-RFE, CMIM, ICAP, SCAD, JMI, CIFE, mRMR, FSCOX, DISR, SNRs, and RankSum, as well as the result with all features, in terms of classification accuracy. The SCES-FS method selected 17 putative miRNAs associated with 10 cancer types and achieved higher classification accuracy than the other methods. Using the 17 selected miRNAs, a web-based multi-class cancer predictor application has been developed.

These selected miRNAs are used in the second part of the proposed method, which employs Cox regression analysis to examine their importance with respect to survival of particular types of cancer. The analysis uses data on expression of the 17 selected miRNAs together with clinical data. A high Cox coefficient value signifies the importance of an miRNA for a particular cancer type. For example, it is found that hsa-mir-375 has the highest Cox coefficient, 0.9882, for the glioblastoma multiform cancer type. Similar results were obtained for other miRNAs, associated with different cancer types. The up- and down-regulations of the 17 selected miRNAs have been computed based on the ANOVA test. Furthermore, network analysis, expression analysis using hierarchical clustering, KEGG pathway analysis, GO enrichment analysis, and PPI network analysis have been performed to assess the biological significance of the selected miRNAs. The network analysis revealed the association of different cancer types with each pair of miRNA and its target mRNA. The hierarchical clustering

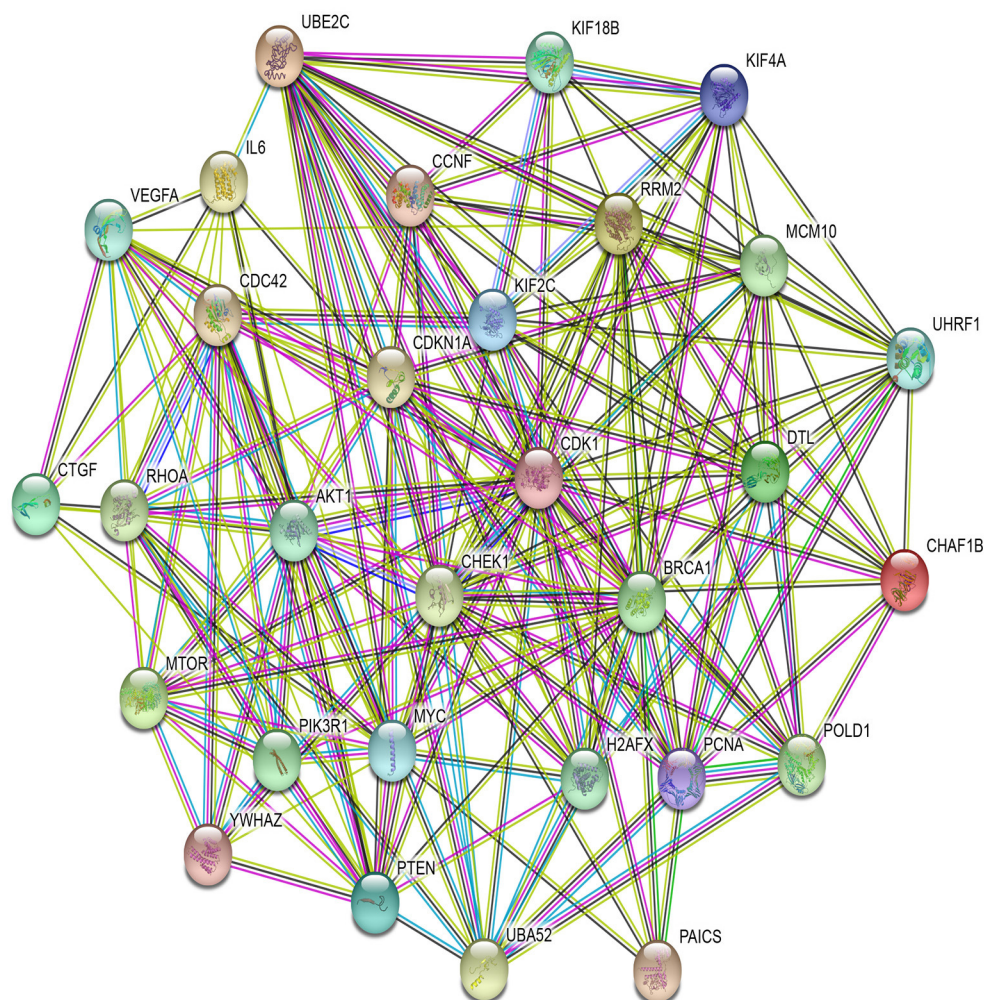


FIGURE 5 | PPI network of top 30 proteins associated with 10 cancer types, with p -value $< 1.0 \times 10^{-16}$ and average node degree 13.7.

analysis demonstrated the effective changes in expression levels of the miRNAs between tumor and normal samples. Both the KEGG and GO enrichment analyses reveals the significant pathways and biological functions in different cancer types. Moreover, using PPI networks, key cancer regulators such as MYC, VEGFA, AKT1, CDKN1A, RHOA, and PTEN are identified. All these evidences suggest that our selected miRNAs play key roles in the development of 10 different types of cancer. A future research direction is the integration of multi-omics data for finding effective regulators pan-cancer biomarkers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://portal.gdc.cancer.gov/> and <http://www.nitttrkol.ac.in/indrajit/projects/mirna-prediction-multicalss/>.

AUTHOR CONTRIBUTIONS

JPS, IS, and AL conceived and designed the experiments. JPS, IS, AL, NG, AD, and PL performed the experiments. JPS, IS, MW, and AD wrote the manuscript. JPS, IS, GB, and DP corrected and edited the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Polish National Science Centre (2014/15/B/ST6/05082, 2019/35/O/ST6/02484), the Foundation for Polish Science (TEAM to DP), and a grant from the Department of Science and Technology, India (Indo-Polish/Polish-Indo project no. DST/INT/POL/P-36/2016). The work was also supported by grant 1U54DK107967-01 Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation within the 4DNucleome NIH

program, and was partially supported by the Academy of Development as the Key to Strengthen Human Resources of the Polish Economy co-financed by the European Union under the European Social Fund. This research was also partially supported under the RENOIR Project of the European Union Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 691152 and by the Ministry of Science and Higher

Education of Poland under grant nos. W34/H2020/2016 and 329025/PnH/2016.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00982/full#supplementary-material>

REFERENCES

- Akhtar, M. M., Micolucci, L., Islam, M. S., Olivieri, F., and Procopio, A. D. (2015). Bioinformatic tools for microRNA dissection. *Nucleic Acids Res.* 44, 24–44. doi: 10.1093/nar/gkv1221
- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* 46, 175–185. doi: 10.1080/00031305.1992.10475879
- Anaissi, A., Goyal, M., Catchpoole, D. R., Braytee, A., and Kennedy, P. J. (2016). Ensemble feature learning of genomic data using support vector machine. *PLoS ONE* 11:e157330. doi: 10.1371/journal.pone.0157330
- Ang, J. C., Mirzal, A., Haron, H., and Hamed, H. N. A. (2016). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 13, 971–989. doi: 10.1109/TCBB.2015.2478454
- Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215–233. doi: 10.1016/j.cell.2009.01.002
- Bennasar, M., Hicks, Y., and Setchi, R. (2015). Feature selection using joint mutual information maximisation. *Expert Syst. Appl.* 42, 8520–8532. doi: 10.1016/j.eswa.2015.07.007
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324
- Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* 30, 89–99. doi: 10.2307/2529620
- Brown, G., Pocock, A., Zhao, M. J., and Lujan, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* 13, 27–66. Available online at: <http://jmlr.org/papers/v13/brown12a.html>
- Bruscella, P., Bottini, S., Baudesson, C., Pawlowsky, J.-M., Feray, C., and Trabucchi, M. (2017). Viruses and miRNAs: more friends than foes. *Front. Microbiol.* 8:824. doi: 10.3389/fmicb.2017.00824
- Cheerla, N., and Gevaert, O. (2017). MicroRNA based pan-cancer diagnosis and treatment recommendation. *BMC Bioinformatics* 18:32. doi: 10.1186/s12859-016-1421-y
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-0881-8
- Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018
- Costa-Silva, J., Domingues, D., and Lopes, F. M. (2017). RNA-Seq differential expression analysis: an extended review and a software tool. *PLoS ONE* 12:e190152. doi: 10.1371/journal.pone.0190152
- Cox, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B* 34, 187–202. doi: 10.1111/j.2517-6161.1972.tb00899.x
- Crow, M., Lim, N., Ballouz, S., Pavlidis, P., and Gillis, J. (2019). Predictability of human differential gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 116, 6491–6500. doi: 10.1073/pnas.1802973116
- Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene* 26, 3279–3290. doi: 10.1038/sj.onc.1210421
- Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 96, 1348–1360. doi: 10.1198/016214501753382273
- Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* 5, 1531–1555. Available online at: <https://www.jmlr.org/papers/v5/fleuret04a.html>
- George, H., and Langley, J. P. (1995). Estimating continuous distributions in Bayesian classifiers. *Proc. Elev. Conf. Uncertain. Artif. Intell.* 69, 338–345.
- Giza, D. E., Vasilescu, C., and Calin, G. A. (2014). Key principles of miRNA involvement in human diseases. *Discoveries* 2:e34. doi: 10.15190/d.2014.26
- Guyon, I., Weston, J., and Barnhill, S. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422. doi: 10.1023/A:1012487302797
- Hansen, N. (2006). The CMA evolution strategy: a comparing review. *Towards New Evol. Comput.* 192, 75–102. doi: 10.1007/11007937_4
- Hansen, N. (2016). The CMA evolution strategy: a tutorial. *Comput. Res. Reposit.* 1–39. Available online at: <http://arxiv.org/abs/1604.00772>
- Hansen, N., Muller, S. D., and Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evol. Comput.* 11, 1–18. doi: 10.1162/106365603321828970
- Hansen, N., and Ostermeier, A. (1996). “Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation,” in *Proceedings of IEEE International Conference on Evolutionary Computation*, 1996 (Nagoya), 312–317. doi: 10.1109/ICEC.1996.542381
- He, W., Zhang, M. G., Wang, X. J., Zhong, S., Shao, Y., Zhu, Y., et al. (2013). KAT5 and KAT6B are in positive regulation on cell proliferation of prostate cancer through PI3K-AKT signaling. *Int. J. Clin. Exp. Pathol.* 6, 2864–2871.
- Hinton, G. E., and Roweis, S. T. (2003). “Stochastic neighbor embedding,” in *Advances in Neural Information Processing Systems*, 857–864. Available online at: <http://papers.nips.cc/paper/2276-stochastic-neighbor-embedding.pdf>
- Hoadley, K., Yau, C., Hinoue, T., Stuart, J. M., Benz, C. C., and Laird, P. W. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Bioinformatics* 173, 291–304. doi: 10.1016/j.cell.2018.03.022
- Hosseinali, N., Aghapour, M., Duijff, P. H., and Baradaran, B. (2018). Treating cancer with microRNA replacement therapy: a literature review. *J. Cell. Physiol.* 233, 5574–5588. doi: 10.1002/jcp.26514
- Huang, H.-Y., Lin, Y.-C.-D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* 48, D148–D154. doi: 10.1093/nar/gkz896
- Jacob, H., Stanisavljevic, L., Storli, K. E., Dahl, K. E. H. O., and Myklebust, M. P. (2017). Identification of a sixteen-microRNA signature as prognostic biomarker for stage II and III colon cancer. *Oncotarget* 8, 87837–87847. doi: 10.18632/oncotarget.21237
- Jacobsen, A., Silber, J., Harinath, G., Huse, J. T., Schultz, N., and Sander, C. (2013). Analysis of microRNA-target interactions across diverse cancer types. *Nat. Struct. Mol. Biol.* 20, 1325–1332. doi: 10.1038/nsmb.2678
- Jakulin, A. (2005). *Machine learning based on attribute interactions* (Ph.D. thesis). Univerza v Ljubljani, Ljubljana, Slovenia.
- Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27
- Kelley, C. T. (1999). Iterative methods for optimization. *Soc. Indus. Appl. Math.* 18, 1–188. doi: 10.1137/1.9781611979020
- Kim, S., Park, T., and Kon, M. (2014). Cancer survival classification using integrated data sets and intermediate information. *Artif. Intell. Med.* 62, 23–31. doi: 10.1016/j.artmed.2014.06.003
- Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

- Latinne, P., Debeir, O., and Decaestecker, C. (2001). "Limiting the number of trees in random forests," in *Proceedings of the Second International Workshop on Multiple Classifier Systems* (Cambridge, UK), 178–187. doi: 10.1007/3-540-48219-9_18
- Li, F., Piao, M., Piao, Y., Li, M., and Ryu, K. H. (2014). A new direction of cancer classification: positive effect of low-ranking MicroRNAs. *Osong Public Health Res. Perspect.* 5, 279–285. doi: 10.1016/j.phrp.2014.08.004
- Li, H., and Yang, B. B. (2014). MicroRNA-in drug resistance. *Oncoscience* 14, 3–4. doi: 10.18632/oncoscience.2
- Li, Y., Wu, F.-X., and Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Brief. Bioinformatics* 19, 325–340. doi: 10.1093/bib/bbw113
- Liang, L., Wei, D. M., Li, J. J., Luo, D. Z., Chen, G., Dang, Y. W., et al. (2018). Prognostic microRNAs and their potential molecular mechanism in pancreatic cancer: a study based on The Cancer Genome Atlas and bioinformatics investigation. *Mol. Med. Rep.* 17, 939–951. doi: 10.3892/mmr.2017.7945
- Lin, E., and Lane, H.-Y. (2017). Machine learning and systems genomics approaches for multi-omics data. *Biomark. Res.* 5:2. doi: 10.1186/s40364-017-0082-y
- Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., et al. (2005). MicroRNA expression profiles classify human cancers. *Nature* 435, 834–838. doi: 10.1038/nature03702
- Ma, J., Dong, C., and Ji, C. (2010). MicroRNA and drug resistance. *Cancer Gene Ther.* 17, 523–531. doi: 10.1038/cgt.2010.18
- McClurg, U. L., and Robson, C. N. (2015). Deubiquitinating enzymes as oncotargets. *Oncotarget* 6, 9657–9668. doi: 10.18632/oncotarget.3922
- Mishra, D., and Sahu, B. (2011). Feature selection for cancer classification: a signal-to-noise ratio approach. *Int. J. Sci. Eng. Res.* 2, 1–7. Available online at: <https://www.ijser.org/viewPaperDetail.aspx?APR1117>
- Mukhopadhyay, A., and Maulik, U. (2013). An SVM-wrapped multiobjective evolutionary feature selection approach for identifying cancer-microRNA markers. *IEEE Trans. Nanobiosci.* 12, 275–281. doi: 10.1109/TNB.2013.2279131
- Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). "How many trees in a random forest?," in *Proceedings of 8th International Conference of Machine Learning and Data Mining in Pattern Recognition* (Berlin), 154–168. doi: 10.1007/978-3-642-31537-4_13
- Paul, P., Chakraborty, A., Sarkar, D., Langthasa, M., Rahman, M., Bari, M., et al. (2018). Interplay between miRNAs and human diseases. *J. Cell. Physiol.* 233, 2007–2018. doi: 10.1002/jcp.25854
- Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159
- Peng, S., Zeng, X., Li, X., Peng, X., and Chen, L. (2009). Multi-class cancer classification through gene expression profiles: microRNA versus mRNA. *J. Genet. Genomics* 36, 409–416. doi: 10.1016/S1673-8527(08)60130-7
- Peng, Y., and Croce, C. M. (2016). The role of MicroRNAs in human cancer. *Signal Transd. Target. Ther.* 1:15004. doi: 10.1038/sigtrans.2015.4
- Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.* 1, 81–106. doi: 10.1007/BF00116251
- Ray, S. S., and Maiti, S. (2015). Noncoding RNAs and their annotation using metagenomics algorithms. *Wiley Interdisc. Rev.* 5, 1–20. doi: 10.1002/widm.1142
- Reimand, J., Wagih, O., and Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* 3:2651. doi: 10.1038/srep02651
- Ros, R., and Hansen, N. (2008). "A simple modification in CMA-ES achieving linear time and space complexity," in *Proceedings of Parallel Problem Solving from Nature* (Dortmund), 296–305. doi: 10.1007/978-3-540-87700-4_30
- Saha, S., Mitra, S., and Yadav, R. K. (2017). A stack-based ensemble framework for detecting cancer MicroRNA biomarkers. *Genom. Proteom. Bioinformatics* 15, 381–388. doi: 10.1016/j.gpb.2016.10.006
- Shrestha, S., Yang, C. D., Hong, H. C., Chou, C. H., Tai, C. S., Chiew, M. Y., et al. (2017). Integrated MicroRNA-mRNA analysis reveals miR-204 inhibits cell proliferation in gastric cancer by targeting CKS1B, CXCL1 and GPRC5A. *Int. J. Mol. Sci.* 19:87. doi: 10.3390/ijms19010087
- Song, C., Zhang, L., Wang, J., Huang, Z., Li, X., Wu, M., et al. (2016). High expression of microRNA-183/182/96 cluster as a prognostic biomarker for breast cancer. *Sci. Rep.* 6:24502. doi: 10.1038/srep24502
- Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). String v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x
- Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., and Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* 18, 1454–1461. doi: 10.1093/bioinformatics/18.11.1454
- Valencia-Sanchez, M. A., Liu, J., Hannon, G. J., and Parke, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev.* 20, 515–524. doi: 10.1101/gad.1399806
- van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426. doi: 10.1016/j.tig.2014.07.001
- Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., et al. (2015). DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.* 43, W460–W466. doi: 10.1093/nar/gkv403
- Waldman, T., Zhang, Y., Dillehay, L., Yu, J., Kinzler, K., Vogelstein, B., et al. (1997). Cell-cycle arrest versus cell death in cancer therapy. *Nat. Med.* 3, 1034–1036. doi: 10.1038/nm0997-1034
- Wang, D., Xin, L., Lin, J. H., Liao, Z., Ji, J. T., Du, T. T., et al. (2017). Identifying miRNA-mRNA regulation network of chronic pancreatitis based on the significant functional expression. *Medicine* 96:e6668. doi: 10.1097/MD.0000000000006668
- Whitney, A. W. (1971). A direct method of nonparametric measurement selection. *IEEE Trans. Comput.* C-20, 1100–1103. doi: 10.1109/T-C.1971.223410
- Wong, N. W., Chen, Y., Chen, S., and Wang, X. (2017). OncomiR: an online resource for exploring pan-cancer microRNA dysregulation. *Bioinformatics* 34, 713–715. doi: 10.1093/bioinformatics/btx627
- Yang, Y., Huang, N., and Kong, L. H. W. (2017). A clustering-based approach for efficient identification of microRNA combinatorial biomarkers. *BMC Genomics* 18:210. doi: 10.1186/s12864-017-3498-8
- Yokoi, A., Yoshioka, Y., Hirakawa, A., Yamamoto, Y., Ishikawa, M., Ikeda, S. I., et al. (2017). A combination of circulating miRNAs for the early detection of ovarian cancer. *Oncotarget* 8, 89811–89823. doi: 10.18632/oncotarget.20688
- Zhang, J., Le, T. D., Liu, L., Liu, B., He, J., Goodall, G. J., et al. (2014). Identifying direct miRNA-mRNA causal regulatory relationships in heterogeneous data. *J. Biomed. Inform.* 52, 438–447. doi: 10.1016/j.jbi.2014.08.005
- Zhang, N., Wang, M., Zhang, P., and Huang, T. (2016). Classification of cancers based on copy number variation landscapes. *Biochim. Biophys. Acta* 1860, 2750–2755. doi: 10.1016/j.bbagen.2016.06.003
- Zhang, P.-W., Chen, L., Huang, T., Zhang, N., Kong, X.-Y., and Cai, Y.-D. (2015). Classifying ten types of major cancers based on reverse phase protein array profiles. *PLoS ONE* 10:e0123147. doi: 10.1371/journal.pone.0123147
- Zhang, Y.-H., Zeng, T., Pan, X., Guo, W., Gan, Z., Zhang, Y., et al. (2019). Screening dys-methylation genes and rules for cancer diagnosis by using the pan-cancer study. *IEEE Access* 8, 489–501. doi: 10.1109/ACCESS.2019.2961402
- Zhou, X., Xu, X., Wang, J., Lin, J., and Chen, W. (2015). Identifying mirna/mrna negative regulation pairs in colorectal cancer. *Sci. Rep.* 5:12995. doi: 10.1038/srep12995

Conflict of Interest: JPS was employed by company Larsen & Toubro Infotech Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Sarkar, Saha, Lancucki, Ghosh, Wlasnowolski, Bokota, Dey, Lipinski and Plewczynski. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Identification of Epigenetic Biomarkers with the use of Gene Expression and DNA Methylation for Breast Cancer Subtypes

Indrajit Saha

Department of Computer
Science and Engineering,
National Institute of Technical
Teachers' Training and Research,
Kolkata, India

Somnath Rakshit

Laboratory of Functional
and Structural Genomics,
Center of New Technologies,
University of Warsaw,
Warsaw, Poland

Michal Wlasnowolski

Faculty of Mathematics
and Information Science,
Warsaw University
of Technology,
Warsaw, Poland

Dariusz Plewczynski

Laboratory of Functional
and Structural Genomics,
Center of New Technologies,
University of Warsaw,
Warsaw, Poland

Abstract—Breast cancer is one of the most deadly cancers. It has four subtypes: *Luminal A (LA)*, *Luminal B (LB)*, *HER2-enriched (HER2-E)* and *Basal-like (BL)*. For the cause of breast cancer subtypes, there are different genetic and epigenetic factors involved in its progression and susceptibility. Thus, the identification of genetic and/or epigenetic biomarkers can be helpful to understand the biological mechanisms better and to improve the diagnostic processes of this disease and its subtypes. Hence, this fact motivated us to investigate the epigenetic factor, such as DNA Methylation, with the integration of gene expression in order to find epigenetic biomarkers for breast cancer subtypes. In this regard, we have identified set of up and down regulated genes for each subtype using differential analysis. Thereafter, regression based feature ranking problem is formed in order to find the DNA Methylation site that is mostly responsible for the change in expression of a gene, which is considered as an epigenetic biomarker. A bagging integrated ensemble of decision trees is used for the same. The results of top ten up and down regulated genes and their corresponding most significant DNA Methylation sites are reported for breast cancer subtypes. Moreover, these genes are validated visually by means of survival and expression plots, showing TF-Gene-DNA Methylation interactions, Protein-Protein interaction network, KEGG pathway and GO enrichment analysis. The results show that top differentially expressed up and down regulated genes viz. MMP11, NUF2, EXO1, HJURP, HOXA4, SYNM, CAV1 and COL4A3BP in breast cancer subtypes may change their expression because of DNA Methylation sites viz. cg22418565, cg26029744, cg24741598, cg04550103, cg25952581, cg02109162, cg18498156 and cg04985097 respectively. The code, datasets and supplementary material are present online¹.

Index Terms—Breast cancer subtypes, Multi-omics data, Differential analysis, DNA Methylation

I. INTRODUCTION

Breast cancer is one of the most frequently diagnosed tumors in women with low survival rate [1]. Due to this fact, the scientific community constantly investigate the mechanism of this cancer in order to improve the diagnosis and treatment procedure. Generally, it is well-known that breast cancer is also associated with multiple DNA mutations and genome

rearrangements which affect cell metabolism [2] like other cancer types. It was observed that in addition to genomic aberrations, level of gene expression is also influenced by epigenetic elements that regulates transcriptional and post-transcriptional activities [3].

Among these epigenetic elements, DNA methylation and histone modifications are having impact on the structure and functionality of chromatin while other epigenetic regulator like microRNA (miRNA) silence the genes by inhibiting or destroying transcripts. [4]. It was also noticed that the potential abnormal epigenetic modifications which can result in aberrant gene expression that may lead to distinct biological and clinical implications [5]. Since DNA Methylation is one of the epigenetic factors, it has a substantial impact on gene expression in response to environmental stimuli [6]. It is found that the iteration of DNA methylation and gene can develop and inhibit the tumor progression [7], and also different kinds of human diseases, like neural [8], [9] or reproductive system diseases [10]. Thus the identification of potential epigenetic biomarkers, i.e. the gene effected due to DNA methylation is crucial and important [11].

The development of high-throughput methods in recent years, such as microarray technologies and RNA-Seq using Next Generation Sequencing (NGS) techniques, for collecting biological data, has enabled the analysis of genome-wide DNA methylation and gene expression profiles [12]. As a result, The Cancer Genome Atlas² (TCGA), using aforementioned techniques on a population scale, a large database has been created. It contains whole-genome methylation and gene expression data of cancer and healthy patients. On the other hand, the analysis of TCGA data using computational methods like classification and regression, allowed to identify potential biomarkers for different types of cancer, such as breast cancer, leukemia, myeloma or lymphomas [3]. As the Breast cancer is a heterogeneous disease, it consists of several subtypes with different molecular and clinical features [13], therefore,

¹<http://www.nittrkol.ac.in/indrajit/projects/epigenetic-mrna-breastcancer-subtypes/>

²<https://tcga-data.nci.nih.gov/tcga/>

finding the epigenetic biomarkers for a particular subtype is challenging [14].

To address the above facts, in this paper, we have identified potential methylation site for a gene that is differentially expressed in healthy and tumor samples of four breast cancer subtypes: *Luminal A* (LA), *Luminal B* (LB), *HER2-enriched* (HER2-E) and *Basal-like* (BL). For this purpose, set of up and down regulated genes for each subtype are identified using differential analysis. Thereafter, bagging integrated ensemble of decision trees is used as regression model in order to rank the features, in this case, DNA Methylation sites for a given gene, that is mostly responsible for the change in expression of a gene, which is considered as an epigenetic biomarker. The results of top ten up and down regulated genes and their corresponding most significant DNA Methylation sites are reported for breast cancer subtypes. The biological analysis shows that these genes are involved in the metabolic pathways of cancer. Therefore, these set of epigenetic biomarkers may consider for further *in vitro* biological analysis in order to investigate the more detailed mechanisms of breast cancer subtype's formation.

II. MATERIAL AND METHOD

A. Dataset preparation

The RNA-Seq data in form of RSEM (RNA-Seq by Expectation Maximization) containing the expression values of gene and DNA Methylation sites are obtained from The Cancer Genome Atlas (TCGA) [15] and later normalized into log2 scale. The dataset contains 294 patients with expression values of 18,303 genes and 407,092 DNA Methylation sites. Furthermore, the breast cancer subtype information is collected from [16]. Only the genes that contains less than 1% zero values are kept while the others are discarded. As a result of this, 14,566, 14,601, 14,488 and 14,753 genes are obtained for LA, LB, HER2-E and BL subtype respectively. Similarly, 332,912, 333,628, 334,259 and 333,494 DNA Methylation sites are obtained for LA, LB, HER2-E and BL subtypes respectively. The number of samples, average age of patients and average follow-up days in each class are mentioned in Table II.

TABLE I
STATISTICS OF THE DATASET

Subtype	ID	Samples	No. of days to last follow up
Luminal A	LA	109	1703.59
Luminal B	LB	47	1554.27
HER2-Enriched	HER2-E	14	1372.85
Basal-Like	BL	41	2007.51
Control	Control	83	1412.72

B. Method

The pipeline of the proposed method to rank and validate the top genes both quantitatively and biologically is shown in Figure 1.

TABLE II
STATISTICS OF TOTAL NUMBER OF GENES AND DNA METHYLATION SITES PRESENT IN THE DATASET

Subtype	Gene		DNA Methylation	
	Total	<1% Zeros	Total	<1% Zeros
LA	18303	14566	407092	332912
LB	18303	14601	407092	333628
HER2-E	18303	14488	407092	334259
BL	18303	14753	407092	333494

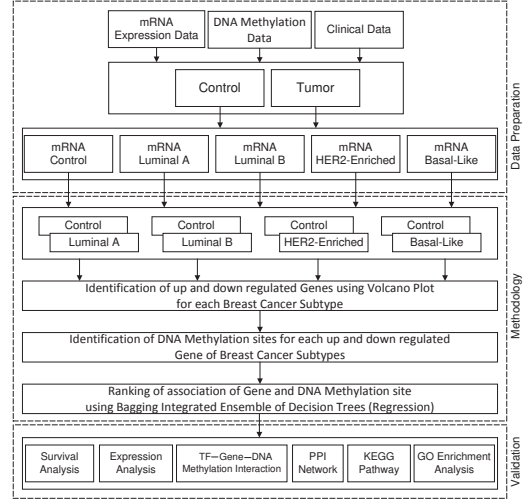


Fig. 1. The pipeline of the experiment to find epigenetic biomarkers for breast cancer subtypes

1) *Identification of Differentially Expressed Genes using Volcano Plot*: For the purpose of identifying up and down regulated genes, volcano plot technique is used. Volcano plot identifies differential genes using the t-test and fold-change (FC) methods. It plots log2 of fold-change value on the X-axis against -log10 of p-value from the t-test on the Y-axis. Genes having positive and negative fold change are called up and down regulated genes respectively. In the present experiment, the up and down regulated genes are obtained using volcano plot for each subtype.

2) *Identification of DNA Methylation sites for each differentially expressed Gene in each breast cancer subtype*: After obtaining the differentially expressed genes, the DNA Methylation sites associated with these genes are found from The Cancer Genome Atlas (TCGA). TCGA contains Gene-DNA Methylation pairs based on the closest gene that is present for

TABLE III
STATISTICS OF UP AND DOWN REGULATED GENES FOR EACH BREAST CANCER SUBTYPE

Subtype	Up	Down
LA	1060	1877
LB	1278	2599
HER2-E	1456	2457
BL	1505	2567

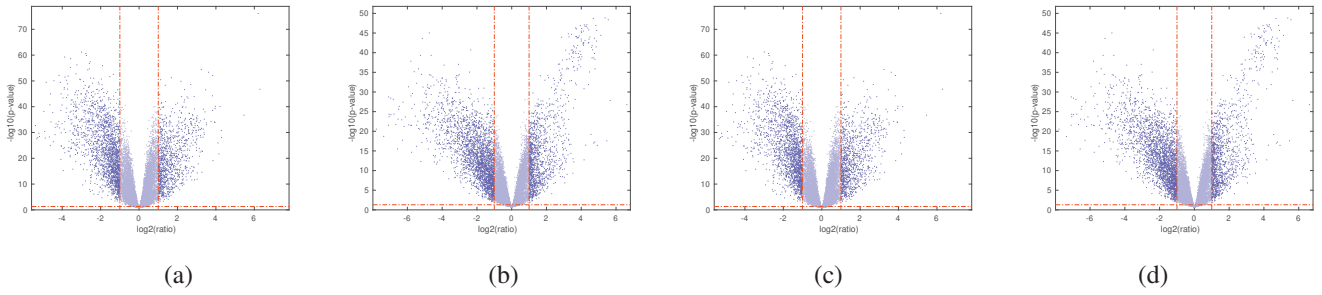


Fig. 2. Volcano plot of gene expression data for patients belonging in breast cancer subtypes: (a) LA, (b) LB, (c) HER2-E and (d) BL and Control

TABLE IV
TOP 10 UP-REGULATED GENES WITH THE LOWEST P-VALUE AND THEIR CORRESPONDING DNA METHYLATION SITE WITH BEST SCORE FOR EACH BREAST CANCER SUBTYPE: LA, LB, HER2-E AND BL

LA				LB			
Gene	Up/Down Regulation provided p-value	DNA Methylation	Rgression Analysis provided Score	Gene	Up/Down Regulation provided p-value	DNA Methylation	Rgression Analysis provided Score
MMP11	7.13E-77	cg22418565	8.24E-02	NUF2	1.98E-49	cg26029744	5.36E-02
HSD17B6	3.77E-55	cg21922731	5.31E-02	NEK2	4.18E-49	cg12820481	1.20E-01
INHBA	4.25E-54	cg21787965	3.74E-02	FOXM1	8.06E-48	cg09976774	4.43E-02
WISP1	7.76E-53	cg04683149	3.55E-02	CEP55	8.31E-48	cg25314624	7.34E-02
RAG1AP1	5.90E-52	cg26189283	1.05E-02	IQGAP3	1.13E-47	cg12617080	5.79E-02
TPM3	3.37E-51	cg03830929	5.01E-03	CENPF	1.30E-47	cg13081150	4.49E-02
NUAK2	1.34E-50	cg20001087	9.68E-03	PLK1	1.75E-47	cg04138181	5.73E-02
BMP8A	2.33E-49	cg11763509	1.97E-02	HJURP	2.33E-47	cg04550103	5.15E-02
LASS2	5.11E-49	cg07422880	6.23E-03	SPAG5	4.97E-47	cg14070845	5.10E-02
COL11A1	1.69E-47	cg03520644	9.62E-02	BUB1	5.15E-47	cg18518914	8.89E-02
HER2-E				BL			
Gene	Up/Down Regulation provided p-value	DNA Methylation	Rgression Analysis provided Score	Gene	Up/Down Regulation provided p-value	DNA Methylation	Rgression Analysis provided Score
EXO1	6.52E-30	cg24741598	3.93E-02	HJURP	5.30E-48	cg04550103	9.28E-02
C16orf59	2.83E-28	cg07326074	2.77E-02	IQGAP3	7.32E-48	cg12441221	6.62E-02
DLGAP5	9.71E-26	cg23678254	8.66E-02	UBE2C	1.67E-47	cg02838589	1.16E-01
NUF2	7.14E-25	cg11823214	3.22E-02	TPX2	3.70E-47	cg09863659	2.24E-01
KIF20A	7.51E-25	cg22106577	5.03E-02	NEK2	4.16E-47	cg17931972	9.42E-02
NUSAP1	7.76E-25	cg25217313	2.83E-02	MELK	4.84E-47	cg14339556	8.92E-02
CDT1	5.86E-22	cg04376887	3.90E-02	KIF2C	4.26E-46	cg11391820F	6.73E-02
HJURP	1.37E-21	cg04550103	5.09E-02	KIFC1	8.32E-46	cg13199639	7.68E-02
FAM111B	1.48E-21	cg23633158	6.33E-02	TROAP	1.63E-45	cg25148733	6.12E-02
CEP55	2.26E-21	cg25314624	8.49E-02	CDC20	1.20E-44	cg16147196	1.02E-01

each DNA Methylation site. Multiple DNA Methylation sites may be present for each gene. Hence, for every gene, a list of DNA Methylation sites are obtained.

3) *Ranking of association of Gene and DNA Methylation site using Bagging Integrated Ensemble of Decision Trees:* The obtained list of DNA Methylation sites are used for the next step of the pipeline. For this purpose, bagging integrated ensemble of decision trees are used. These sites are ranked based on the score obtained from using bagging integrated ensemble of decision trees.

III. EXPERIMENTAL RESULTS

This section describes the experimental setup and the obtained results.

A. Experimental Testbed

The volcano plots and bagging integrated ensemble of decision trees have been implemented in Matlab R2017a while other computations have been done using Pandas 0.24 and Numpy 1.14 in Python 3.6.5. An Intel i5 processor with 4

cores and 8 GB RAM has been used for all computational purposes.

B. Results

1) *Differentially expressed Genes:* Using volcano plots, differentially expressed genes are obtained for all subtypes. For this purpose, the fold change value and the p-value are calculated. All genes having fold change value >2 and p-value <0.05 are termed as differentially expressed genes. These genes are further named as up (positive fold change) or down (negative fold change) regulated genes. The number of such genes is shown in Table III. The volcano plots are shown in Figure 2.

2) *Ranking of DNA Methylation sites for the top 10 Genes from each subtype:* The DNA Methylation sites that are obtained from TCGA are ranked based on bagging integrated ensemble of decision trees. The ranked lists are reported in Tables IV and V respectively for the down and up regulated genes. It is seen that many genes that are well known for breast cancer such as HOXA4 [17], SYNM [18], MMP11 [19], etc.

TABLE V

TOP 10 DOWN-REGULATED GENES WITH THE LOWEST P-VALUE AND THEIR CORRESPONDING DNA METHYLATION SITE WITH BEST SCORE FOR EACH BREAST CANCER SUBTYPE: LA, LB, HER2-E AND BL

LA				LB			
Gene	Up/Down Regulation provided p-value	DNA Methylation	Rgression Analysis provided Score	Gene	Up/Down Regulation provided p-value	DNA Methylation	Rgression Analysis provided Score
HOXA4	2.39E-61	cg25952581	1.70E-02	SYNM	2.32E-44	cg02109162	9.28E-02
SPRY2	1.41E-60	cg00185066	2.46E-02	PPP1R12B	1.95E-41	cg24376793	1.48E-02
RYR3	5.53E-60	cg08642292	6.16E-02	RYR3	4.69E-40	cg08642292	6.95E-02
TMEM220	3.50E-58	cg02025583	3.17E-02	TTYH1	4.74E-40	cg21287054	1.23E-01
NDRG2	1.68E-57	cg18081258	1.99E-02	CRYAB	1.47E-38	cg12947833	1.44E-01
ADAMTS5	6.06E-57	cg12078031	1.86E-02	GSN	9.09E-38	cg14399183	4.34E-02
IL11RA	1.25E-56	cg21504624	1.62E-02	KCNMB1	1.47E-37	cg16425489	8.96E-02
PAMR1	6.18E-56	cg20027133	2.61E-02	HSPB6	1.27E-36	cg13558754	1.18E-01
LMOD1	2.01E-55	cg26914267	1.52E-02	LYVE1	7.04E-36	cg18343862	1.21E-01
ANXA1	5.36E-55	cg21222681	2.88E-02	LTBP4	8.70E-36	cg03354707	2.19E-02
HER2-E				BL			
Gene	Up/Down Regulation provided p-value	DNA Methylation	Rgression Analysis provided Score	Gene	Up/Down Regulation provided p-value	DNA Methylation	Rgression Analysis provided Score
CAV1	4.52E-24	cg18498156	2.25E-02	COL4A3BP	6.59E-30	cg04985097	8.63E-03
AOC3	2.37E-21	cg11744144	6.86E-02	C1QTNF7	7.81E-30	cg03290977	9.87E-02
EPHB1	4.65E-21	cg15000071	3.65E-02	SLC25A18	1.87E-29	cg18003231	5.51E-02
HSPB2	6.89E-21	cg00514609	1.72E-02	KCNIP2	2.75E-29	cg07447260	8.58E-02
GPAM	3.85E-19	cg16794749	1.11E-01	LOC728264	3.29E-29	cg17736336	6.61E-02
KLHL31	5.13E-19	cg13840445	7.16E-02	RDH5	3.44E-29	cg21156320	1.20E-01
CRHBP	8.75E-19	cg01071966	7.54E-02	GSN	5.07E-29	cg14417974	2.54E-02
SORBS1	2.18E-18	cg03417182	4.73E-02	FAM13A	6.07E-29	cg18006637	3.65E-02
MT1M	1.32E-17	cg05925949	4.12E-02	LEPR	2.01E-28	cg26342890	3.81E-02
RNF150	5.32E-17	cg25898520	4.60E-02	TNS1	2.02E-28	cg05386769	4.38E-02

are obtained from the proposed pipeline. Apart from this, the top DNA Methylation site affecting each of these genes and the weight obtained from bagging integrated ensemble of decision trees is also shown in the tables.

TABLE VI
INTERACTION OF TOP GENES WITH DNA METHYLATION AND TRANSCRIPTION FACTOR (TF)

TF	Gene	DNA Methylation
SIRT1	ADAMTS5	cg12078031
GATA6	CAV1	cg18498156
PHF8	CDC20	cg16147196
TSG101	CEP55	cg25314624
CEBPZ	COL11A1	cg03520644
ESR1	CRHBP	cg01071966
TP53	CRYAB	cg12947833
TCF3	EXO1	cg24741598
FLI1	FOXM1	cg09976774
TFAP2A	HOXA4	cg25952581
HSF1	HSD17B6	cg21922731
E2F1	KIF2C	ch.1.1391820F
GATA1	LTBP4	cg03354707
SP1	MMP11	cg22418565
WT1	NDRG2	cg18081258
E2F4	PLK1	cg04138181
CREB1	SPRY2	cg00185066
MED1	UBE2C	cg02838589

3) *Survival Analysis using Kaplan-Meier Plots:* Kaplan-Meier Analysis is a way of measuring the fraction of subjects living for a certain amount of time after treatment. The obtained genes are used for Kaplan-Meier Analysis. It is seen from the plots that the obtained genes are associated with the survival of a patient in a significant manner. The KM plots along with the boxplots of expression of selected genes from the obtained list of top genes are shown in Figure 3. Here, the boxplots show the change of expression of genes for a

TABLE VII
KEGG PATHWAY ANALYSIS SHOWING THE COMMON PATHWAYS FOR MULTIPLE SUBTYPES

ID	LA	LB	HER2-E	BL
hsa05200	✓	✓	✓	✓
hsa04110	✓	✓	✓	✓
hsa04151	✓	✓	✓	✓
hsa04350	✓	✓	✓	✓
hsa04068	✓	✓	✓	✓
hsa04919	✓	✓	✓	✓
hsa04010	✓	✓	✓	✓
hsa04115	✓	✓	✓	✓
hsa04620	✓	✓	✓	✓
hsa04668	✓	✓	✓	✓
ID	Term			
hsa05200	Pathways in cancer			
hsa04110	Cell cycle			
hsa04151	PI3K-Akt signaling pathway			
hsa04350	TGF-beta signaling pathway			
hsa04068	FoxO signaling pathway			
hsa04919	Thyroid hormone signaling pathway			
hsa04010	MAPK signaling pathway			
hsa04115	p53 signaling pathway			
hsa04620	Toll-like receptor signaling pathway			
hsa04668	TNF signaling pathway			

population are correlating with the KM plots, e.g., for MMP11 gene, the population of patients that has high expression is having low survival rate. Vice versa results are also obtained for down regulated genes.

4) *TF-Gene-DNA Methylation interaction:* A TF-Gene-DNA Methylation table is constructed using the top genes and their corresponding top DNA Methylation sites. Additionally, TRRUST database is used to identify the TFs that are associated with these genes. This table is shown in Table VI. It is observed from the table that many TFs that are well known for breast cancer such as TP53, ESR1, GATA1, etc are present for the obtained genes.

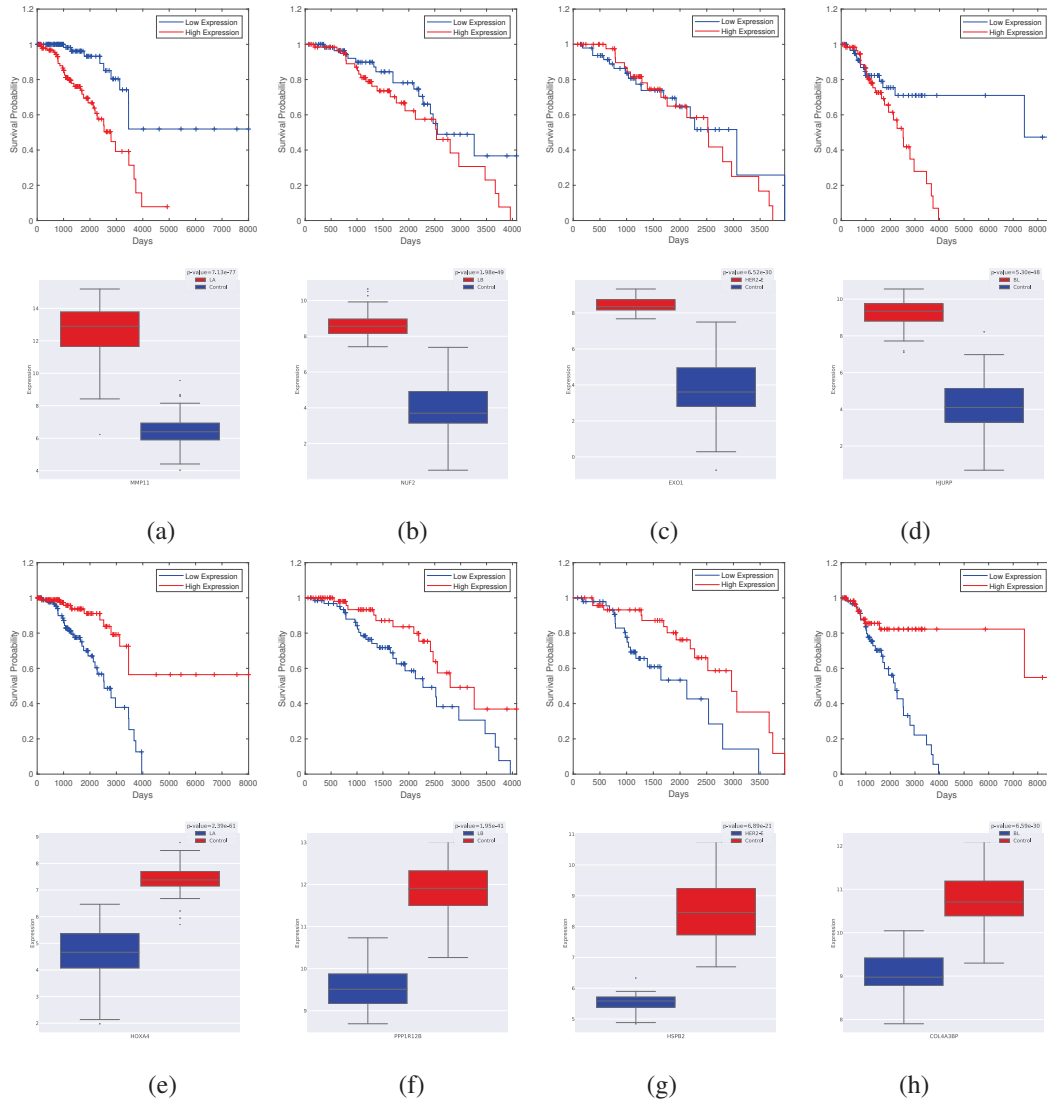


Fig. 3. Kaplan-Meier plots and boxplots of expression of the top subtype specific genes in (a)-(d): Up and (e)-(h): Down regulated genes for LA, LB, HER2-E and BL subtypes respectively. Here, blue color represents low expression and red color represents high expression.

5) *Protein-Protein Interaction Networks*: The obtained top genes are used to find their associated TFs from TRRUST database. Afterwards, the TFs that are present in three or more subtypes are considered to prepare a protein-protein interaction (PPI) network. This is done using the STRING database. The obtained PPI Network is shown in Figure 4. Apart from this, a bar plot showing the degree of the top ten TFs are shown in the same figure. It is seen from the figure that TFs that are known to be influential in breast cancer such as TP53, ESR1, GATA1, etc. are the TFs with the highest degree.

6) *KEGG Pathway Analysis*: KEGG Pathway analysis has been performed using the Enrichr tool by using the top ten up and down regulated genes from each subtype. It is seen that many well known pathways that are related to breast cancer are obtained. Some examples include Pathways in cancer, PI3K-Akt signaling pathway, TGF-beta signaling pathway, etc. This

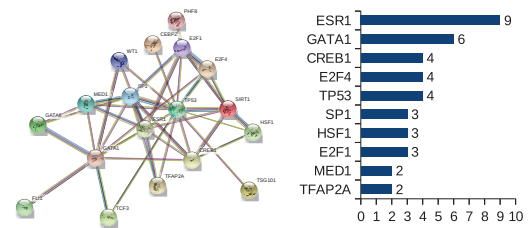


Fig. 4. PPI Network consisting of TFs that are common to three or more subtypes. A bar plot showing the degree of the top 10 nodes in the PPI Network is also present.

suggests that the genes that are obtained through the proposed pipeline are significantly related to breast cancer. Table VII shows the subtypes which contain the significant pathways.

TABLE VIII
GENE ENRICHMENT ANALYSIS SHOWING THE COMMON PATHWAYS
ACROSS MULTIPLE SUBTYPES

ID	LA	LB	HER2-E	BL
GO:0045892	✓	✓	✓	✓
GO:0030182	✓	✓	✓	✓
GO:0010628	✓	✓	✓	✓
GO:0051726	✓	✓	✓	✓
GO:0007265	✓	✓	✓	✓
GO:0042981	✓	✓	✓	✓
GO:2000648	✓	✓	✓	✓
GO:0010629	✓	✓	✓	✓
GO:0043066	✓	✓	✓	✓
GO:0043069	✓	✓	✓	✓
ID	Term			
GO:0045892	Negative regulation of transcription, DNA-templated			
GO:0030182	Neuron differentiation			
GO:0010628	Positive regulation of gene expression			
GO:0051726	Regulation of cell cycle			
GO:0007265	Ras protein signal transduction			
GO:0042981	Regulation of apoptotic process			
GO:2000648	Positive regulation of stem cell proliferation			
GO:0010629	Negative regulation of gene expression			
GO:0043066	Negative regulation of apoptotic process			
GO:0043069	Negative regulation of programmed cell death			

7) *GO Enrichment Analysis*: GO Enrichment Analysis for Biological Process is also performed using Enrichr tool. For this purpose, the top ten up and down regulated genes are used as input. Significant terms like Positive regulation of gene expression, Regulation of cell cycle, Ras protein signal transduction, etc. are obtained. Table VIII shows the subtypes which contain the significant terms.

IV. CONCLUSION

Epigenetic biomarkers such as genes that are effected by DNA Methylation sites are a promising field of research, that can provide better understanding and diagnostics to breast cancer therapy. Since breast cancer is heterogenous in nature, identification of its epigenetic biomarkers should be performed by considering the subtypes. In this work, we have focused on finding specific DNA Methylation site which has impact on a gene transcription level. In our research, we have used the TCGA database containing genome-wide DNA methylation and gene expression data on population scale. As part of the differential analysis, we have found list of significant up and down regulated genes, while regression analysis provides rank of DNA methylation sites for a given gene. Our results have been verified using survival plot, PPI network analysis and KEGG and GO enrichment analysis, which indicate the significance of top epigenetic biomarkers as they are belonging in cancer metabolic pathways. Therefore, these set of epigenetic biomarkers can be considered for *in vitro* analysis in order to investigate the mechanism of breast cancer subtypes in more detail.

ACKNOWLEDGEMENTS

This work has been supported by the Polish National Science Centre (2014/15/B/ST6/05082), Foundation for Polish

Science (TEAM to DP) and by the grant from the Department of Science and Technology, India under Indo-Polish/Polish-Indo project No.: DST/INT/POL/P-36/2016. Moreover, the work was co-supported by grant 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation" within 4DNucleome NIH program.

REFERENCES

- [1] R. Segal, K. Miller, and A. Jemal, "Cancer statistics, 2018," *CA: A Cancer Journal for Clinicians*, vol. 68, pp. 7–30, 2018.
- [2] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, p. 719, 2009.
- [3] A. Nebbioso, F. P. Tambaro, C. Dell'Aversana, and L. Altucci, "Cancer epigenetics: moving forward," *PLoS Genetics*, vol. 14, no. 6, p. e1007362, 2018.
- [4] C. Carlberg and F. Molnár, *Human Epigenomics*. Springer, 2018.
- [5] M. Shivakumar, Y. Lee, L. Bang, T. Garg, K.-A. Sohn, and D. Kim, "Identification of epigenetic interactions between miRNA and DNA methylation associated with gene expression as potential prognostic markers in bladder cancer," *BMC Medical Genomics*, vol. 10, no. 1, p. 30, 2017.
- [6] K. D. Robertson, "DNA methylation and human disease," *Nature Reviews Genetics*, vol. 6, no. 8, p. 597, 2005.
- [7] X. Hao, H. Luo, M. Krawczyk, W. Wei, W. Wang, J. Wang, K. Flagg, J. Hou, H. Zhang, S. Yi *et al.*, "DNA methylation markers for diagnosis and prognosis of common cancers," *Proceedings of the National Academy of Sciences*, vol. 114, no. 28, pp. 7414–7419, 2017.
- [8] J. N. Kuehner, E. C. Bruggeman, Z. Wen, and B. Yao, "Epigenetic Regulations in Neuropsychiatric Disorders," *Frontiers in Genetics*, vol. 10, 2019.
- [9] E. L. Crowgey, A. G. Marsh, K. G. Robinson, S. K. Yeager, and R. E. Akins, "Epigenetic machine learning: utilizing DNA methylation patterns to predict spastic cerebral palsy," *BMC Bioinformatics*, vol. 19, no. 1, p. 225, 2018.
- [10] Z. Li, X. Zhuang, J. Zeng, and C.-M. Tzeng, "Integrated analysis of DNA methylation and mRNA expression profiles to identify key genes in severe oligozoospermia," *Frontiers in Physiology*, vol. 8, p. 261, 2017.
- [11] L. T. Kagohara, G. L. Stein-O'Brien, D. Kelley, E. Flam, H. C. Wick, L. V. Danilova, H. Easwaran, A. V. Favorov, J. Qian, D. A. Gaykalova *et al.*, "Epigenetic regulation of gene expression in cancer: techniques, resources and analysis," *Briefings in Functional Genomics*, vol. 17, no. 1, pp. 49–63, 2017.
- [12] K. Wang and X. Liu, "Integrative analysis of genome-wide expression and methylation data," *Journal of Biometrics and Biostatistics*, vol. 4, pp. 4–6, 2013.
- [13] X. Dai, L. Xiang, T. Li, and Z. Bai, "Cancer hallmarks, biomarkers and breast cancer molecular subtypes," *Journal of Cancer*, vol. 7, no. 10, p. 1281, 2016.
- [14] D. C. Temian, L. A. Pop, A. I. Irimie, and I. Berindan-Neagoe, "The Epigenetics of Triple-Negative and Basal-Like Breast Cancer: Current Knowledge," *Journal of Breast Cancer*, vol. 21, no. 3, pp. 233–243, 2018.
- [15] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.
- [16] C. G. A. Network *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.
- [17] M. Mustafa, J.-Y. Lee, and M. H. Kim, "CTCF negatively regulates HOXA10 expression in breast cancer cells," *Biochemical and Biophysical Research Communications*, vol. 467, no. 4, pp. 828–834, 2015.
- [18] E. Noetzel, M. Rose, E. Sevinc, R. Hilgers, A. Hartmann, A. Naami, R. Knüchel, and E. Dahl, "Intermediate filament dynamics and breast cancer: aberrant promoter methylation of the synemin gene is associated with early tumor relapse," *Oncogene*, vol. 29, no. 34, p. 4814, 2010.
- [19] R. G. de Vega, D. Clases, M. L. Fernández-Sánchez, N. Eiró, L. O. González, F. J. Vizoso, P. A. Doble, and A. Sanz-Medel, "Mmp-11 as a biomarker for metastatic breast cancer by immunohistochemical-assisted imaging mass spectrometry," *Analytical and Bioanalytical Chemistry*, vol. 411, no. 3, pp. 639–646, 2019.

Gliwice, 12.08.2023

Recenzja rozprawy doktorskiej mgr inż. Michała Własnowolskiego
pt. *Computational Modelling and Analysis of the Three-Dimensional Structure*
of Human Genome at the Population Scale

Ukończonej na Wydziale Matematyki i Nauk Informacyjnych
Uniwersytetu Warszawskiego

Pod opieką promotora prof. dr hab. Dariusza Plewczyńskiego

Zawartość pracy

Trójwymiarowa architektura chromatyny w jądrze komórkowym jest ważnym czynnikiem regulacji i funkcjonowania genomu. Zaawansowane technologie mapowania całego genomu takie jak Hi-C czy ChIA-PET pozwalają eksperymentalnie badać interakcje chromatyny i dostarczają danych, które można wykorzystać do rekonstrukcji trójwymiarowej struktury genomu. W szczególności wynikiem tych wysokoprzepustowych technologii są dane w postaci częstotliwości interakcji parami między loci genomowymi, które muszą zostać przekształcone we względne odległości fizyczne. Ponieważ te techniki eksperymentalne dostarczają danych na temat kontaktów chromatynowych, a nie bezpośrednich informacji strukturalnych, istnieje potrzeba tworzenia algorytmów oraz narzędzi obliczeniowych potrafiących wykorzystać informacje na temat częstotliwości kontaktów w celu zamodelowania struktury 3D. Możliwość modelowania struktury trójwymiarowej chromatyny jest szczególnie ważna w przypadku próby zrozumienia jak zmienność strukturalna związana z występowaniem delecji, insercji, duplikacji oraz inwersji wpływa na przestrzenną organizację genomu, co ma w konsekwencji wpływ na mechanizmy regulacji transkrypcji genów. W tym kontekście szczególnie interesująca jest możliwość analizy przestrzennych interakcji chromatyny związanych z regionami *enhancerów* i *promotorów* genów, które odgrywają kluczową rolę w regulacji transkrypcji genów.

Przedstawiona praca doktorska została złożona w postaci zbioru opublikowanych i powiązanych tematycznie artykułów naukowych. Na cykl publikacji składają się cztery artykuły naukowe, trzy z nich zostały opublikowane w bardzo dobrych czasopismach naukowych, czwarty znajduje się w recenzji. W dwóch opublikowanych artykułach i w trzecim, który jest w recenzji, doktorant jest pierwszym autorem.

Pierwszy z artykułów [P1] pt. „*Spatial chromatin architecture alteration by structural variations in human genomes at the population scale*” opublikowany został w czasopiśmie *Genome Biology* (IF 18.01, 200 pkt MNiSW). W pracy tej przeanalizowano po raz pierwszy w skali populacyjnej częstość występowania zmienności wariantów strukturalnych i ich wpływ na organizację chromatyny oraz zmienność jej topologii. W celu stworzenia modeli 3D chromatyny wykorzystana jest autorska metoda 3D-GNOME, która modyfikuje referencyjne interakcje genomowe i domeny topologiczne uzyskane na podstawie danych eksperymentalnych wprowadzając informacje o zmienności strukturalnej wynikającej ze zmienności genetycznej. W pracy umieszczono też szeroką populacyjną analizę wariantów strukturalnych.

Drugi artykuł [P2] składający się na cykl publikacji pt. *3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome* opublikowany został w czasopiśmie *Nucleic Acid Research* (IF=19.16, 200 pkt. MNiSW), sekcja Web Server Issue. W pracy tej przedstawiono narzędzie implementujące algorytm 3D-GNOME w wersji 2.0. Algorytm ten szczegółowo został opisany w pracy P1 a w połączeniu z opublikowanym narzędziem umożliwia użytkownikom przewidywanie zmian w trójwymiarowej konformacji chromatyny dla wybranego regionu ludzkiego genomu. Użytkownik może również przesłać własną listę wariantów strukturalnych w celu ich zamodelowania.

Kolejny artykuł [P3] składający się na cykl publikacji opublikowany został w formie nierecenzowanego preprintu na serwerach bioRxiv oraz został przesłany do recenzji w czasopiśmie *Bioinformatics* (IF 6.931, 200 pkt MNiSW). Jest to praca zatytułowana „*cudaMMC - GPU-enhanced Multiscale Monte Carlo Chromatin 3D Modelling*”. W pracy tej przedstawiono modyfikację algorytmu 3D-GNOME w celu przyspieszenia obliczeń. Przyspieszenie uzyskano poprzez zrównoleglenie obliczeń z wykorzystaniem architektury GPU.

Ostatni, czwarty artykuł [P4] składający się na cykl publikacji, pt. *3D-GNOME 3.0: a three-dimensional genome modelling engine for analysing changes of promoter-enhancer contacts in the human genome* opublikowany został w czasopiśmie *Nucleic Acid Research* (IF=19.16, 200 pkt MNiSW), sekcja Web Server Issue. Praca ta przedstawia kolejną wersję narzędzia 3D-GNOME. Narzędzie rozszerzone zostało o nowy, znacznie większy zbiór danych interakcji chromatynowych, a także zaktualizowane warianty strukturalne do najnowszej wersji genomu ludzkiego GRCh38. Istniejący silnik zastąpiono algorytmem cudaMMC oraz zintegrowano serwis z klastrem obliczeniowym co umożliwiło do 25x przyspieszenie obliczeń. Serwis został też wzbogacony o narzędzia do wizualizacji oraz porównywania wyników.

Opinia o rozprawie

Należy podkreślić, iż problem, który podjąć się rozwiązać Doktorant jest niewątpliwie ważny i wpisujący się w trendy najnowszych badań w dziedzinie. Rozwój algorytmów oraz narzędzi do przewidywania oraz wizualizacji struktury trójwymiarowej chromatyny jest pierwszym krokiem do zrozumienia w jaki sposób zmiany struktury przestrzennej chromatyny wpływają na mechanizmy regulacji transkrypcji genów. Analiza zróżnicowania populacyjnego wariantów strukturalnych jest ważna w kontekście poznania patogennego potencjału wariantów strukturalnych zmieniających organizację chromatyny wyższego rzędu.

Przedstawione w ramach pracy doktorskiej artykuły stanowią niewątpliwie spójny cykl powiązanych ze sobą publikacji w tematyce rozprawy doktorskiej.

Dobór artykułów i kolejność prezentacji wyników badań jest spójny i logiczny. Pierwsza praca zawiera analizę częstości występowania zmian strukturalnych chromatyny w populacji i motywację do wprowadza narzędzie 3D-GNOME. Następne publikacje przedstawiają kolejne ulepszone wersje algorytmu. Opublikowane z sukcesem wyniki badań publikacje świadczą o tym, iż Doktorant opanował warsztat badawczy i potrafi realizować badania naukowe.

Spis literatury dobrze oddaje aktualny stan wiedzy w zakresie, którego dotyczy rozprawa. Wiele z cytowanych prac to pozycje najnowsze co świadczy o śledzeniu przez Doktoranta na bieżąco literatury przedmiotu.

Przedstawiony w rozprawie cel, który brzmi:

Opracowanie, wdrożenie i współtworzenie narzędzi obliczeniowych służących do generowania i analizy trójwymiarowych modeli chromatyny oraz badania potencjalnego wpływu struktury przestrzennej chromatyny na aktywność genetyczną komórek.

jest jasno sformułowany, a przedstawiony cykl artykułów świadczy o tym, że Doktorantowi udało się go osiągnąć.

Uwagi krytyczne i dyskusyjne

Na początku tej części chciałbym podkreślić, że nie znalazłem w przedstawionych wynikach żadnych zasadniczych błędów merytorycznych czy niewłaściwych rozumowań. Wszystkie poniższe uwagi wynikają z chęci podjęcia dyskusji i dialogu na temat niektórych aspektów pracy. Uwagi te nie obniżają mojej pozytywnej oceny pracy.

Charakter publikacji w czasopismach Nucleic Acid Research Web Server Issue, powoduje, że skupiają się one bardziej na przedstawieniu cech narzędzia niż analizie metody. Po pierwsze, chciałabym zaznaczyć, że tego typu publikacje są bardzo ważne, gdyż udostępniają daną metodę szerokiemu środowisku naukowemu

przyczyniając się do powstawania nowych wyników i wniosków nie tylko w grupie badawczej, w której została stworzona metoda. Niemniej jednak pewną wadą tych publikacji jest brak pogłębionych analiz wyników udostępnionych metod. Nie jest więc jasne czy dla nowych wersji algorytmu Doktorant wykonał analizy podobne do analiz przedstawionych w pracy [P1]? Czy wprowadzenie nowych modyfikacji do algorytmu - poza przyspieszeniem obliczeń i aktualizacją do nowej wersji genomu - pozwoliło uzyskać dodatkową wiedzę biologiczną lub nowe wnioski odnośnie wpływu wariantów strukturalnych na strukturę przestrzenną genomu w analizie populacyjnej?

Niedosyt budzi też niezbyt konkretne przedstawienie przez Doktoranta wkładu własnego w powstanie publikacji. W przypadku pracy doktorskiej przedstawionej jako cykl powiązanych tematycznie publikacji jest to niezwykle istotne, ponieważ pozwala ocenić jaka część z opublikowanych prac stanowi oryginalny wkład Doktoranta. Załączono co prawda oświadczenia współautorów, niemniej są one bardzo ogólne (przykładowo jest to koordynacja zespołu, konceptualizacja i metodologia – wspólnie z Promotorem) lub obejmują zadania mocno techniczne takie jak integracja zbioru danych czy formatów danych. Przedstawione publikacje są wieloautorskie, co w przypadku tak złożonych metod i narzędzi jest normalną praktyką, jednak Doktorant powinien w jasny sposób zaakcentować swój twórczy wkład w powstanie publikacji, a w szczególności w rozwój kolejnych wersji algorytmu 3D-GNOME.

Wnioski

Pan mgr inż. Michała Własnowolski przedstawił zbiór opublikowanych i powiązanych tematycznie artykułów naukowych stanowiących oryginalne rozwiązanie problemu naukowego z zakresu trójwymiarowego modelowania chromatyny w kontekście analizy wariantów strukturalnych. Należy podkreślić, iż opracowane przez Doktoranta metody udostępnione zostały środowisku naukowemu w postaci serwisów internetowych.

Poza publikacjami składającymi się na pracę doktorską, Doktorant jest również współautorem trzech publikacji opublikowanych w bardzo dobrych czasopismach naukowych (odpowiednio 100, 140 oraz 100 punktów MNiSW) oraz jednej publikacji konferencyjnej.

Biorąc pod uwagę powyższą ocenę, stwierdzam, że przedstawiona do oceny praca doktorska w pełni odpowiada warunkom określonym w Art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (tekst jednolity Dz. U. z 2023 r. poz. 742 z późn. zm.) i na tej podstawie wnoszę do Wysockiej Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o dopuszczenie mgr inż. Michała Własnowolskiego do dalszych etapów przewodu doktorskiego.



dr inż. hab. Aleksandra Gruca



Wrocław, 8.08.2023

Recenzja rozprawy doktorskiej Pana mgra Michała Własnowolskiego pt.
"Computational Modelling and Analysis of the Three-Dimensional Structure of Human
Genome at the Population Scale"

Odczyt informacji genetycznej zwłaszcza u organizmów eukariotycznych, a szczególności u zwierząt, jest procesem bardzo skomplikowanym i wciąż słabo poznanym. Wymaga on nie tylko samej sekwencji genomowego DNA oraz oddziaływania czynników regulacyjnych i epigenetycznych, ale również odpowiedniej przestrzennej organizacji chromatyny. Aby zrozumieć te zjawiska stosowane są techniki eksperymentalne, jak ChIA-PET i Hi-C, dostarczające danych o przestrzennych kontaktach w chromatynie. Jednakże są one czasochłonne i kosztowne, dlatego nie można ich stosować na szeroką skalę, np. w badaniach populacyjnych, które są niezbędne do poznania różnic i mechanizmów molekularnych wielu chorób. W związku z tym istnieje potrzeba rozwijania metod komputerowych do badania i interpretowania struktury i oddziaływań chromatyny.

Dlatego bardzo słusznie ambitnym przedmiotem pracy doktorskiej Pana mgra Michała Własnowolskiego stało się opracowanie zaawansowanych narzędzi informatycznych służących do modyfikacji wzorów kontaktów chromatynowych wynikających ze zmian sekwencji DNA. Dzięki temu możliwe jest generowanie i porównywanie różnych modeli przestrzennych chromatyny na poziomie populacyjnym.

Rozprawa doktorska została napisana w języku angielskim ze streszczeniem w języku polskim. Zawiera ona: spis treści, wykaz rycin, wstęp, główny rozdział przedstawiający najważniejsze osiągnięcia naukowe związane z rozprawą doktorską i przedstawione w załączonych czterech pracach, rozdział zawierający pozostałe osiągnięcia doktoranta, wnioski i przyszłe badania, bibliografię, kopie publikacji będące przedmiotem rozprawy, oświadczenia współautorów i kopie dodatkowych publikacji.

Wstęp stanowi zwięzłe, ale dobre wprowadzenie do idei określania struktury przestrzennej chromatyny. Przedstawiono w nim metody eksperymentalne (Hi-C i ChIA-PET) i sposób prezentowania oddziaływań w postaci diagramów łukowych. Najwięcej miejsca poświęcono podejściu obliczeniowemu 3D-GENOME, które w oparciu o eksperymenty ChIA-PET przedstawia trójwymiarowy model chromatyny. Zakłada ono interakcje między miejscami chromatyny, w których pośredniczą białka RNAPII, CTCF i kohezyny. Podejście to oparte o symulacje metodą wyżarzania metodą Monte Carlo i zakładające hierarchiczną organizację chromatyny zostało wykorzystane w doktoracie. W wystarczający sposób scharakteryzowano także regulację ekspresji genów na poziomie epigenetycznym poprzez metylację cytozyny i modyfikacje histonów oraz wiązanie się białek do chromatyny i elementów regulatorowych w DNA, tj. promotorów, wyciszaczy i wzmacniaczy. Wyjaśniono, że ich interakcje w przestrzeni trójwymiarowej wpływają na odpowiednią ekspresję informacji genetycznej, a zaburzenia w ich oddziaływaniu mogą być związane z chorobami.

W osobnym rozdziale jasno sformułowano cele pracy doktorskiej, którymi było opracowanie i zastosowanie narzędzi obliczeniowych służących do generowania i analizy trójwymiarowego modelu chromatyny oraz zbadania wpływu przestrzennej struktury chromatyny na aktywność genetyczną. Cele pracy są poprawnie sformułowane i wszystkie zostały właściwie zrealizowane.

Osiągnięcia związane z rozprawą doktorską zostały przedstawione w czterech pracach. W trzech z nich doktorant jest pierwszym autorem, a w jednej czwartym. Wyniki trzech prac zostały opublikowane w prestiżowych czasopismach (*Genome biology* i *Nucleic Acids Research*), a czwarta jest w trakcie recenzji w również prestiżowym czasopiśmie *Bioinformatics*. Wyniki zaprezentowane w tych pracach są spójne i układają się w jedną całość.

Załączone oświadczenia autorów nie pozostawiają wątpliwości, że Pan Michał Własnowolski miał wiodący i istotny wpływ na koncepcję prac i uzyskanie oraz opisanie wyników w publikacjach będących przedmiotem pracy doktorskiej. Przedstawiona rozprawa opisująca wyniki prac jest oryginalnym wkładem doktoranta.

W pierwszej publikacji doktorant przedstawił wpływ wariantów genetycznych na strukturę trójwymiarową chromatyny i opisał narzędzie do przewidywania zmiany tej struktury w oparciu o warianty strukturalne (SV, structural variants) w genomie człowieka, takie jak delecje, duplikacje, insercje i inwersje. Są one związane z wieloma chorobami i wadami rozwojowymi. Opracowane narzędzie opiera się o dane ChIA-PET w połączeniu z wariantami

strukturalnymi genomów uzyskanymi w ramach 1000 Genomes Consortium. Opracowane narzędzie umożliwia badania w skali populacyjnej, co jest istotne, aby powiązać zmiany genomowe z fenotypem w skali globalnej. Warto podkreślić, że opracowane narzędzie i badania są pierwszymi, które wiążą zmienność struktury przestrzennej genomu człowieka w skali całej populacji. W pracy tej określono tendencję wariantów strukturalnych do gromadzenia się w przestrzennie oddziałujących segmentach genomowych. Pokazano, że zróżnicowana transkrypcja genów jest ściśle związana ze zmiennością sieci interakcji chromatyny, w których pośredniczy polimeraza RNA II. Wykazano również, że interakcje, w których bierze udział białko CTCF, są konserwatywne w genomach człowieka, ale zawierają dużo polimorfizmów pojedynczych nukleotydów (SNP) związanych z chorobami. Ponadto stwierdzono, że granice domen topologicznych w chromatynie są stosunkowo częstymi miejscami duplikacji, co sugeruje, że tego typu zmiany mogą być ważnym mechanizmem wpływającym na przestrzenną organizację genomu.

W drugiej publikacji opisano integrację wcześniej utworzonego narzędzia z serwisem internetowym 3D-GNOME, który posiada przyjazny interfejs dla użytkownika oraz użyteczne narzędzia umożliwiające analizę różnic w trójwymiarowej strukturze chromatyny dla różnych wariantów strukturalnych. W początkowej wersji, narzędzie to generowało trójwymiarowe modele chromatyny oparte na danych z metody Chromatin Conformation Capture (3C). Serwis dostarcza również narzędzia do wizualizacji i analizy struktury przestrzennej. Użytkownicy mogą przewidzieć strukturę chromatyny dla wybranego regionu genomu człowieka określając jego współrzędne. Poza wariantami strukturalnymi uzyskanymi z 1000 Genomes Project można również podawać własne warianty. Podobnie, poza danymi dotyczącymi interakcji białek CTCF i RNAPII z chromatyną w oparciu o eksperymenty ChIA-PET dla linii limfoblastycznej komórek GM12878 można podawać własne dane tego typu. Usługa internetowa wykorzystuje framework webowy Flask z zapytaniami do bazy danych MySQL. Dane są przetwarzane w oparciu o skrypty Pythonie, natomiast główne oprogramowanie jest napisane w C++, PHP i R. Serwis 3D-GNOME generuje diagramy kontaktowe chromatyny, wykresy rozkładu długości miejsc kontaktowych, podsumowujące statystyki i modele przestrzenne chromatyny. Modele można oglądać poprzez interaktywną przeglądarkę zaimplementowaną w WebGL. Wygenerowane modele można pobrać w standardowych formatach i oglądać w innych programach.

W trzeciej publikacji Pan Michał Własnowolski zaprezentował narzędzie cudaMMC, czyli nową wersję 3D-GNOME, które wykorzystując kart graficzne (GPU) do akceleracji

obliczeniowej, potrafi 25-krotnie przyspieszyć obliczenia przewidywania struktury przestrzennej w porównaniu z wersją oryginalną. Jest to istotne podejście ze względu na intensywny przyrost liczby danych z eksperymentów ChiA-PET i Hi-C. Dzięki temu można przeprowadzać wiele przewidywań struktury chromatyny w krótkim czasie, co jest istotne dla analiz statystycznych w oparciu o dane populacyjne. Takie analizy są kluczowe dla zrozumienia wpływ trójwymiarowej struktury chromatyny na regulację transkrypcji i badanie zaburzeń w tej strukturze z powodu zmiany odległości między elementami regulatorowymi, takimi jak wzmacniacze i promotory. Zadanie to było dużym wyzwaniem, ponieważ w przypadku symulowanego wyżarzania metodą Monte Carlo trudno jest zoptymalizować obliczenia równoległe i sekwencyjne na GPU i CPU, gdyż energia globalna dla całej struktury chromatyny jest obliczana po poprawieniu lokalnych oddziaływań. Wymagało to napisania nowego oprogramowania. Obliczenia oparte o GPU charakteryzują się także większą stabilnością.

W czwartej publikacji opisano aktualizację serwisu internetowego 3D-GNOME. Udoskonalenie polegało na uwzględnieniu w modelach trójwymiarowych chromatyny narzędzia do analizy zmian odległości między wzmacniaczami i promotorami biorącymi udział w regulacji transkrypcji genów. W badaniach tych zastosowano większy zbiór danych interakcji chromatyny ChIA-PET o lepszej rozdzielczości oraz uaktualnione informacje o wariantach strukturalnych z danych genomowych o pokryciu 30-krotnym. Dodatkowo w tej aktualizacji narzędzie cudaMMC zintegrowano z serwerem WWW, a obliczenia optymalizujące proces generowania modeli 3D przerzucono na klaster obliczeniowy EdenN, który wykorzystuje karty graficzne Nvidia DGX A100. To znacznie przyspiesza obliczenia. W nowym narzędziu, po podaniu danych dotyczących wzorów interakcji chromatyny oraz wymodelowaniu struktury przestrzennej chromatyny, a także zmapowaniu promotorów i wzmacniaczy genów porównywane są rozkłady odległości między nimi dla sekwencji referencyjnej i zmienionego wariantu strukturalnego. Wyniki podawane są w formie tabelarycznej, wykresów łukowych, modeli przestrzennych i wykresów pudełkowych. Zmiany odległości między promotorami i wzmacniaczami mogą mieć istotne znaczenie w zmianie ekspresji genów. Dlatego to narzędzie jest bardzo pomocne i ważne, aby przewidywać takie zmiany.

Część odpowiadająca wnioskowi jest napisana przejrzysto i zawierają najważniejsze informacje uzyskane z przeprowadzonych badań i opracowanego narzędzia. Interesujące jest nawiązanie do przyszłych analiz dotyczących paleogenomów człowieka, które mogą wnieść wiele ciekawych informacji na temat zmiany ekspresji informacji genetycznej w ewolucji.

Doktorant włożył dużo trudu w opracowanie nowatorskiego narzędzia i przeprowadzone analizy, a przedstawione opisy wyników świadczą o dużej dojrzałości naukowej doktoranta i umiejętności wydobywania najważniejszych informacji z uzyskanych rezultatów. Nie mam zastrzeżeń do metodyki przeprowadzonych analiz. Opisy są dobrze przedstawione pod względem formalnym. Praca i artykuły są napisane poprawnym językiem i stylem.

Mam tylko drobne uwagi. W Streszczeniu polskim zamiast meatzoa powinno się użyć terminu zwierząt albo po prostu Metazoa z dużej litery lub Animalia. Abstract angielski nie odpowiada dokładnie Streszczeniu w języku polskim. Na przykład nie ma w nim informacji, że zastosowanie kart graficznych zwiększyło obliczenia do 25 razy. W opisie polskim zamiast "enhancerów" dałbym "wzmacniaczy". Zdanie: "In the case of metazoans, such as humans, the complete genetic information in each cell is identical (apart from mutations accumulated during development)." można zastosować dla wszelkich organizmów wielokomórkowych, nie tylko zwierząt. Na Fig. 2.4 w części Output jest napisane "gene – enhancer distances". Czy nie powinno być "promotor-enhancer distances"? W nowej wersji programu przy określaniu istotności różnic między rozkładami odległości między promotorami i wzmacniaczami zastosowano test Mann–Whitneya U. Test ten adekwatny, jeśli dane nie mają rozkładu normalnego. Nie zawsze musi tak być i w tym przypadku, gdyby dane miały taki rozkład, można by stosować test t-Studenta. Można to rozważyć w uaktualnionej wersji. Mam jeszcze pytanie: dlaczego narzędzie nazwano 3D-GNOME, a nie 3D-GENOME?

Stwierdzone przeze mnie zastrzeżenia nie rzutują jednak na bardzo pozytywną ocenę pracy, a przedstawione powyżej uwagi nie zmniejszają wartości ocenianej rozprawy. Tematyka pracy doktorskiej jest bardzo zasadna, ponieważ istnieje potrzeba tworzenia narzędzi i analizy struktury przestrzennej chromatyny w aspekcie całej populacji człowieka z uwzględnieniem zmian strukturalnych i elementów regulujących transkrypcję, co stało się przedmiotem rozprawy. Należy podkreślić, że opracowany program jest zaawansowany, została polepszona szybkość i skuteczność jego działania, a analizy zostały przeprowadzone skrupulatnie na dużym zbiorze danych dostarczając nowych informacji na temat ekspresji informacji genetycznej. Reasumując chciałbym stwierdzić, że recenzowana rozprawa z artykułami stanowi istotny wkład w przewidywanie i analizę struktury przestrzennej chromatyny.

Na uwagę zasługuje dodatkowy dorobek publikacyjny doktoranta obejmujący cztery prace, uczestnictwo w czterech projektach badawczych oraz trzy wizyty naukowe. Świadczą one o wszechstronności i dojrzałości doktoranta.

Uważam, więc, że przedstawiona do recenzji rozprawa doktorska spełnia wszystkie wymogi Ustawy o Stopniach Naukowych. Zgłaszam, zatem wniosek do Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o uznanie rozprawy Pana mgr Michała Własnowolskiego za odpowiadającą wymogom stawianym rozprawom doktorskim i o dopuszczenie doktoranta do dalszych etapów przewodu doktorskiego. W związku z tym, że doktorant miał postawiony trudny cel badawczy i go efektywnie rozwiązał, a wyniki zostały przedstawione w czterech bardzo dobrych pracach proponuję wyróżnić rozprawę.



Prof. dr hab. Paweł Mackiewicz