Metody obliczeniowe w analizie rearanżacji chromosomowych

Autoreferat

Barbara Poszewiecka

Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

Wstęp

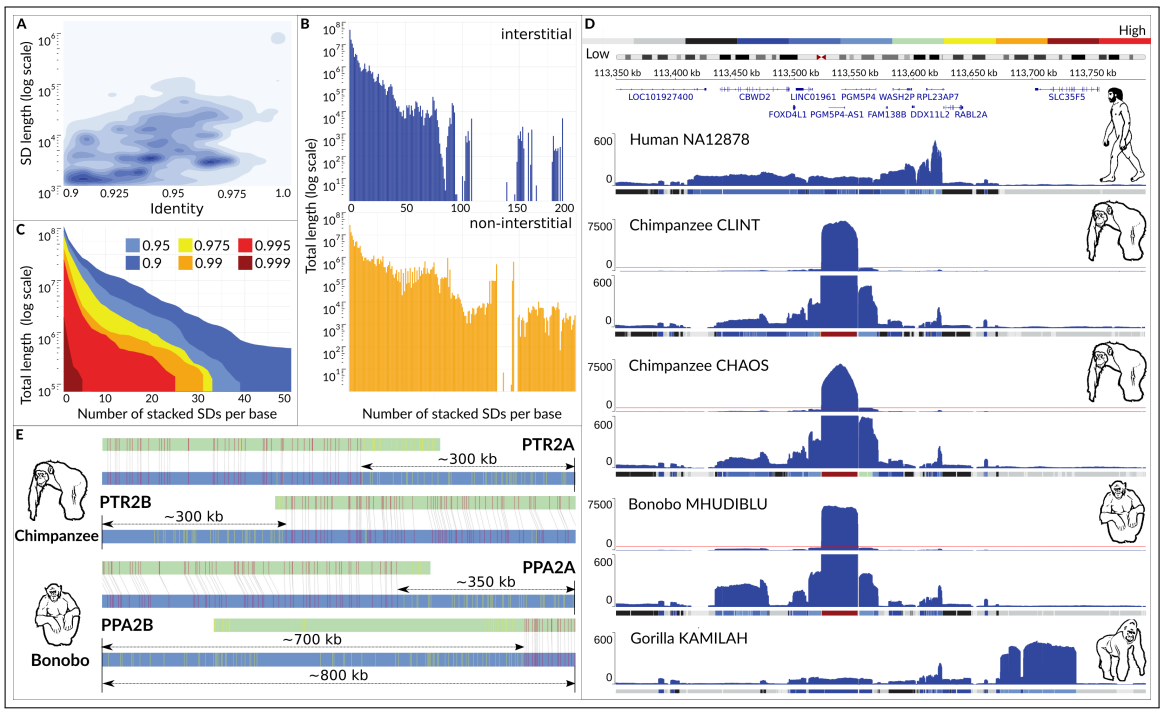
W ciągu ostatnich dziesięcioleci, wiedza na temat biologii komórki została zrewolucjonizowana przez wprowadzenie nowych metod biotechnologicznych. W szczególności fascynujące możliwości w naukach przyrodniczych zostały stworzone przez zastosowanie wysokoprzepustowych technologii sekwencjonowania, znanych również jako sekwencjonowanie nowej generacji (NGS). Metody te umożliwiają równoległe sekwencjonowanie wielu cząsteczek DNA lub RNA pochodzących z całego genomu, w szybki, ekonomiczny i powtarzalny sposób. Ciągły postęp w technologiach związanych z technologiami NGS umożliwia ich wykorzystanie w różnych dziedzinach biologii molekularnej, w tym w genomice (wyjaśnianiającej strukturę, funkcję i ewolucję genomu), transkryptomice (zajmującej się analizą transkryptów RNA) i epigenomice (badającej zmiany fenotypowe, które nie są powodowane przez zmiany w genomie). Aby w pełni wykorzystać ich potencjał we wszystkich tych dyscyplinach, potrzebne są nowe metody obliczeniowe zdolne do przetwarzania ogromnych ilości danych oraz umożliwiające wyciąganie z nich poprawnych i znaczących biologicznych konkluzji.

Główne wyniki

Wszystkie wyniki przedstawione w rozprawie doktorskiej dotyczą problemu wykrywania i interpretacji zmian w architekturze genomu kształtowanej przez rearanżacje chromosomowe. Każdy z przedstawionych w niej rozdziałów opisuje nowe metody bioinformatyczne i ich zastosowanie w rozwiązywaniu istotnych zagadnień biologicznych lub klinicznych. Wyniki zaprezentowane w pracy są zdecydowanie interdyscyplinarne, a każda metoda jest zilustrowana znaczącym studium przypadku z wykorzystaniem rzeczywistych danych biomedycznych.

Asemblacja rejonów wzbogaconych w segmentalne duplikacje

Coraz pełniejsze zrozumienie architektury i ewolucji genomu jest możliwe dzięki szybkiemu rozwojowi technologii sekwencjonowania, zarówno jeśli chodzi o dokładność, jak i długość generowanych odczytów, oraz ciągłemu postępowi w obliczeniowych metodach ich przetwarzania (Huddleston *et al.*, 2014; il Sohn and Nam, 2016; Amarasinghe *et al.*, 2020). Rozwój w tych dziedzinach potwierdził fakt, że segmentalne duplikacje (fragmenty genomu określone jako dłuższe niż 1000 par zasad powielone z identycznością większą niż 90%) odgrywają kluczową rolę w kształtowaniu architektury genomu oraz umożliwiają zduplikowanym genom proces przysposabiania się do nowych funkcji, wpływając znacząco na ewolucję genomu, szczególnie jeśli chodzi o genomy małp naczelnych (patrz Rysunek 1). Należy podkreślić fakt, że segmentalne duplikacje stanowią znaczną części genomu człowieka, oszacowaną przez Konsorcjum Telomere-to-Telomere jako 7% (Vollger *et al.*, 2022).

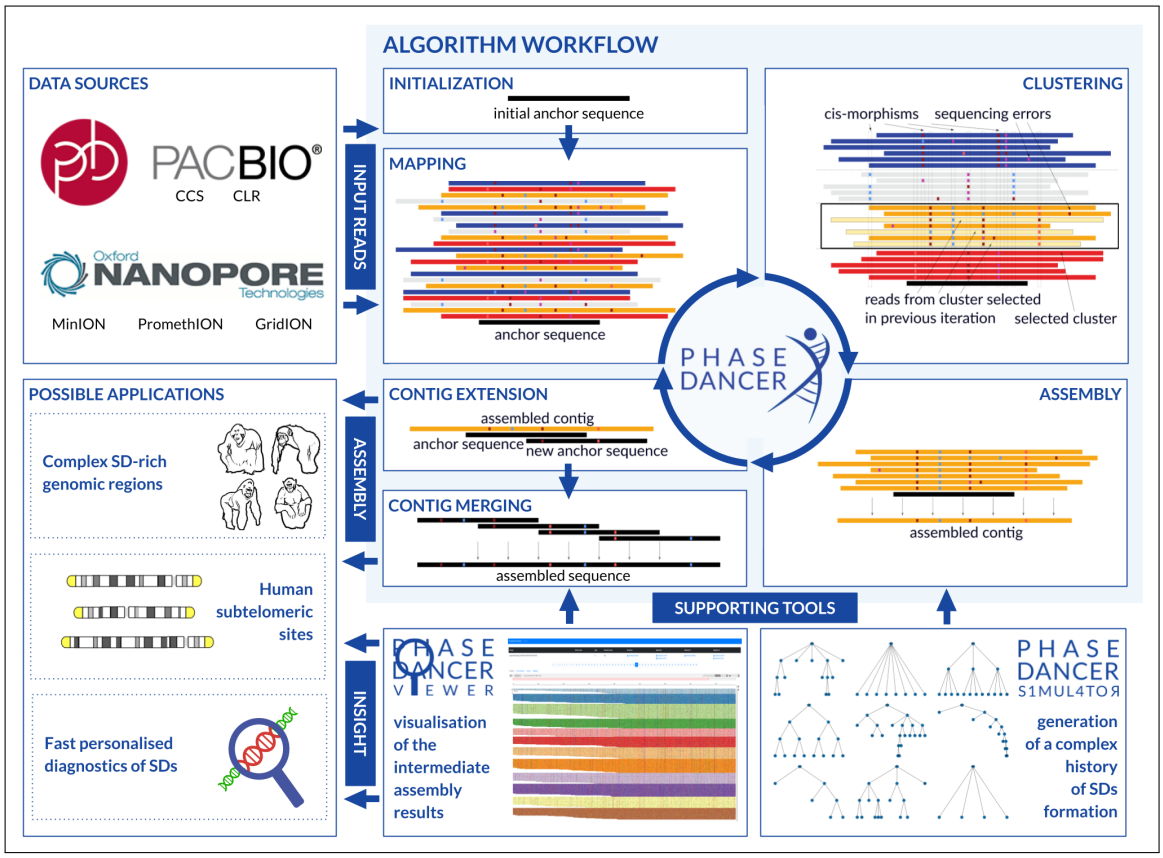


Rysunek 1: Charakterystyka segmentalnych duplikacji i motywacja badań. Analizy dotyczące genomu człowieka powstały na podstawie najnowszej wersji genomu referencyjnego człowieka dostarczonego przez konsorcjum T2T: (**A**) Wykres konturowy przedstawiający liczność segmentalnych duplikacji o danej identyczności pomiędzy sekwencjami (90–100%, oś x) i całkowitej długości (Mb, oś y, skala logarytmiczna), wysycenie koloru niebieskiego odpowiada liczbie segmentalnych duplikacji. (**B**) Wykres słupkowy przedstawiający całkowitą długość segmentalnych duplikacji (oś x) o danej liczbie kopii występujących w interstycjalnych (top, blue) i nieinterstycjalnych (bottom, yellow) regionach genomu. (**C**) Wykres powierzchniowy przedstawiający całkowitą długość segmentalnych duplikacji (Mb, skala logarytmiczna, oś y) o liczbie kopii większej lub równej wartości zaznaczonej na osi x i minimalnym procencie identyczności odpowiadającemu kolorowi powierzchni wykresu. (**D**) Znormalizowany histogram głębokości pokrycia odczytów pochodzących z sekwencjonowania w technologii PacBio CCS dla człowieka (NA12878), dwóch osobników szympansa (Clint, Chaos), bonobo (Mhudilbu) i goryla (Kamilah) na podstawie uliniowienia do regionu w pobliżu miejsca fuzji na chromosomie drugim człowieka. W przypadku bonobo i obu szympansów pokazano dwa wykresy prezentujące głębokości pokrycia. Górny wykres prezentuje wszystkie dane, dolny ma ograniczoną wartość na osi y i tym samym pozwala na na zaprezentowanie danych z wyłączeniem obszaru o wyjątkowo wysokim pokryciu. Czerwona linia na każdym z górnych wykresów wskazuje granicę osi y dolnego wykresu. (**E**) Dane z technologii optical genome mapping zostały użyte do zbadania kompletności subtelomerowych regionów w genomach szympansa i bonobo (panTro5, panTro6 i panPan3). Oszacowano, że na każdym z subtelomerów brakuje co najmniej 0,3 Mb sekwencji DNA.

Asemblacja *de novo* regionów wzbogaconych w segmentalne duplikacje jest wymagającym zadaniem obliczeniowym, głównie z powodu dużej liczby błędów w danych pochodzących z sekwencjonowania długimi odczytami. Istniejącym asemblerom często nie udaje się złożyć fragmentów genomu, które są wzbogacone w segmentalne duplikacje, zwłaszcza gdy użyte odczyty są znacząco krótsze niż pochodzące z technologii Ultra-Long Oxford Nanopore lub gdy są mniejszej dokładności niż dane PacBio CCS (Vollger *et al.*, 2019b). Możliwym podejściem pozwalającym na wykorzystanie takich danych do asemblacji segmentalnych duplikacji jest stworzenie narzędzia sprofilowanego do tego celu.

W ramach rozprawy doktorskiej opracowane zostało narzędzie PhaseDancer – asembler, który w swoim działaniu wykorzystuje innowacyjne podejście polegające na lokalnym składaniu regionów wzbogaconych w segmentalne duplikacje przy pomocy danych z sekwencjonowania długimi odczytami. W przeciwieństwie do istniejących narzędzi, które wykorzystują podejście *top-down* operując jednocześnie na wszystkich zgromadzonych danych, PhaseDancer generuje wyniki stopniowo, zgodnie z paradygmatem *bottom-up*, wydłużając sekwencję początkową przy użyciu odczytów o wystarczającym podobieństwie. Algorytm przyjmuje jako dane wejściowe początkową, krótką sekwencję kotwicy, którą jest rozszerzana iteracyjnie poprzez powtarzanie czterech głównych kroków. Najpierw odczyty są mapowane na sekwencję kotwicy przy użyciu indeksu wszystkich odczytów przechowywanych w pamięci RAM. W drugim kroku, zmapowane odczyty są klastrowane przy użyciu randomizowanej procedury i wybierany jest klaster, który dzieli najwięcej odczytów z poprzednio wybranym. Trzeci krok to złożenie wybranych odczytów w kontig, a czwarty to rozszerzenie aktualnej sekwencji kotwicy przy pomocy złożonego kontigu do nowej sekwencji kotwicy, która będzie przetwarzana w następnej iteracji. Integracja najnowocześniejszych komponentów użytych podczas implementacji poszczególnych kroków działania PhaseDancera pozwoliła na generowanie kontigów z fragmentów powtórzonych kilkadziesiąt razy z podobieństwem nie przekraczającym 99.9% (patrz Rysunek 2).

Narzędzie PhaseDancer zostało dodatkowo wyposażone w przeglądarkę klastrów oraz symulator segmentalnych duplikacji. Narzędzie PhaseDancerViewer wizualizuje każdą iteracje algorytmu poprzez przedstawienie przeglądarki genomowej pokazującej klastry odczytów zmapowane na sekwencję kotwicy. Aplikacja ta pomaga w dostrojeniu parametrów asemblacji, korzystaniu z asemblera w trybie półnadzorowanym, kontroli poprawności asemblacji oraz umożliwia stawianie hipotez biologicznych dotyczących struktury klastrów. PhaseDancerSimulator generuje *in silico* sekwencje segmentalnych duplikacji na podstawie zdefiniowanych przez użytkownika scenariuszy ich ewolucyjnej historii oraz zestawu odpowiednich parametrów określających charakterystykę generowanych sztucznych odczytów.



Rysunek 2: Schemat działania algorytmu PhaseDancer i opis towarzyszących mu narzędzi. PhaseDancer operuje na danych z sekwencjonowania długimi odczytami (Oxford Nanopore, PacBio). Proces asemblacji polega na rozszerzaniu krótkiej sekwencji kotwicy, w którym iterowane są cztery główne kroki: (i) mapowanie odczytów na sekwencję kotwicy, (ii) klastrowanie zmapowanych odczytów i wybór klastra najbardziej pasującego do poprzednio wybranego, (iii) składanie wybranych odczytów w kontig, oraz (iv) rozszerzanie bieżącej sekwencji kotwicy przy użyciu kontigu do nowej sekwencji kotwicy przetwarzanej w następnej iteracji. Po wykonaniu ustalonej ilości iteracji algorytm łączy wszystkie sekwencje kotwic w wynikowy kontig. Do asemblera dołączone są narzędzia pomocnicze: PhaseDancerViewer pozwalający na wizualizację klastrów uzyskanych w każdej iteracji oraz PhaseDancerSimulator, który generuje scenariusze segmentalnych duplikacji i odpowiadające im odczyty pozwalajac na dogłębną walidację narzędzi składających segmentalne duplikacje.

W celu walidacji narzędzia PhaseDancer, zastosowany został zestaw klonów BAC wygenerowanych z ludzkiej haploidalnej linii komórkowej używany standardowo do testowania asemblerów pod kątem poprawności składania segmentalnych duplikacji oraz sztuczne segmentalne duplikacje wygenerowane przez symulator. Sekwencje złożonych przez PhaseDancer kontigów odpowiadającym klonom BAC zostały porównane z wynikami asemblacji wygenerowanymi przez narzędzia Flye i Wtdgb2 przy użyciu danych z sekwencjonowania długimi odczytami w technologii PacBio z 45-krotnym pokryciem. Liczba złożonych przez PhaseDancer klonów BAC zdecydowanie przewyższyła asemblery ogólnego przeznaczenia, rozwiązując 292 z 341 klonów (85,5%), podczas gdy asemblery Flye i Wtdgb2 rozwiązały odpowiednio 91 (26,69%) i 77 (22,58%) klonów BAC.

Poprawność działania narzędzia zweryfikowano przy pomocy scenariuszy ewolucyjnych wygenerowanych przez PhaseDancerSimulator porównując je do wyników asemblacji kilku powszechnie używanych asemblerów działających na danych z sekwencjonowa-

nia długimi odczytami (Canu (Koren *et al.*, 2017), Wtdbg2 (Ruan and Li, 2020), Flye (Kolmogorov *et al.*, 2019), Miniasm (Li, 2016), and SDA (Vollger *et al.*, 2019a)).

Z wygenerowanych 10 scenariuszy ewolucyjnych, z liczbą kopii w zakresie od 2 do 12, PhaseDancer rozwiązał wszystkie symulowane segmentalne duplikacje uzyskując identyczność sekwencji o wartości Phred powyżej 29 (ponad 99,8%), podczas gdy innym asemblerom ogólnego przeznaczenia udało się rozwiązać maksymalnie jedną segmentalną duplikację w każdym z przypadków, a jedynie asembler SDA dedykowany do asemblacji segmentalnych duplikacji rozwiązał scenariusze z liczbą kopii nie przekracającą czterech.

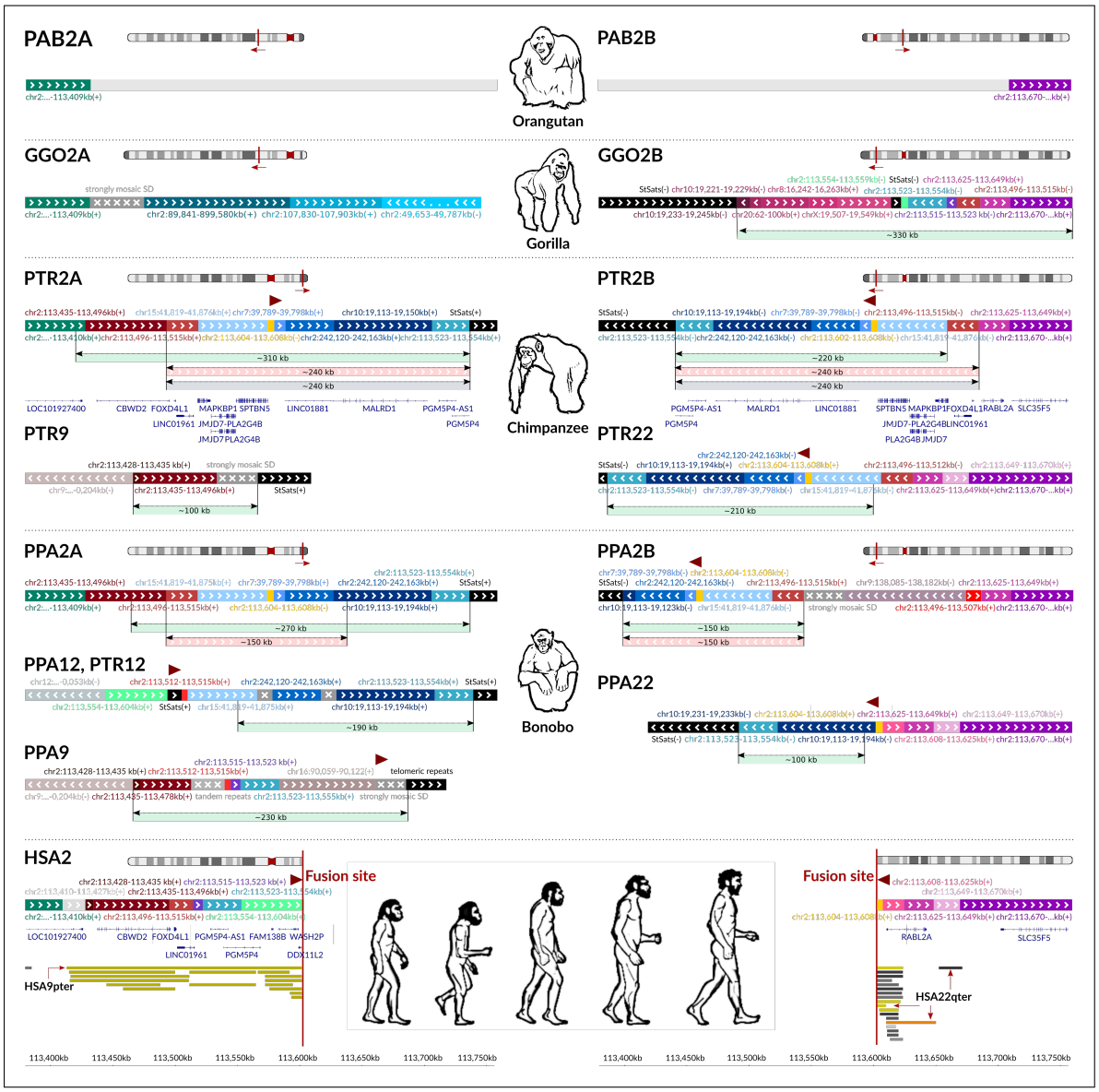
Ze względu na poprawność wyników walidacji, algorytm mógł zostać zastosowany do wiarygodnej asemblacji regionów subtelomerowych wybranych chromosomów małp człekokształtnych (szympansa, bonobo, goryla i orangutana) syntenicznych do miejsca fuzji na chromosomie 2 człowieka (patrz Rysunek 3). Miało to na celu wyjaśnienie mechanizmu redukcji liczby chromosomów podczas ewolucji praczłowieka po oddzieleniu się od wspólnego przodka człowieka, szympansa i bonobo. Rozszerzenie sekwencji referencyjnych pozwoliło na zaproponowanie modelu fuzji dwóch chromosomów, w wyniku której liczba chromosomów u człowieka spadła z 48 do 46, porównaniu z małpami człekokształtnymi. Podjęto również próbę uzasadnienia hipotezy, mówiącej że fuzja stanowiła prawdopodobną przewagę ewolucyjną, która mogła ułatwić jej utrwalenie i akumulację w populacji (patrz Rysunek 4).

Asembler PhaseDancer oraz proponowany model fuzji prowadzący do powstania chromosomu 2 człowieka został zaprezentowany na konferencjach *26th Annual International Conference on Research in Computational Molecular Biology* oraz *American Society of Human Genetics 2022 Annual Meeting* podczas sesji posterowej.

Metody datowania dużych zdarzeń ewolucyjnych

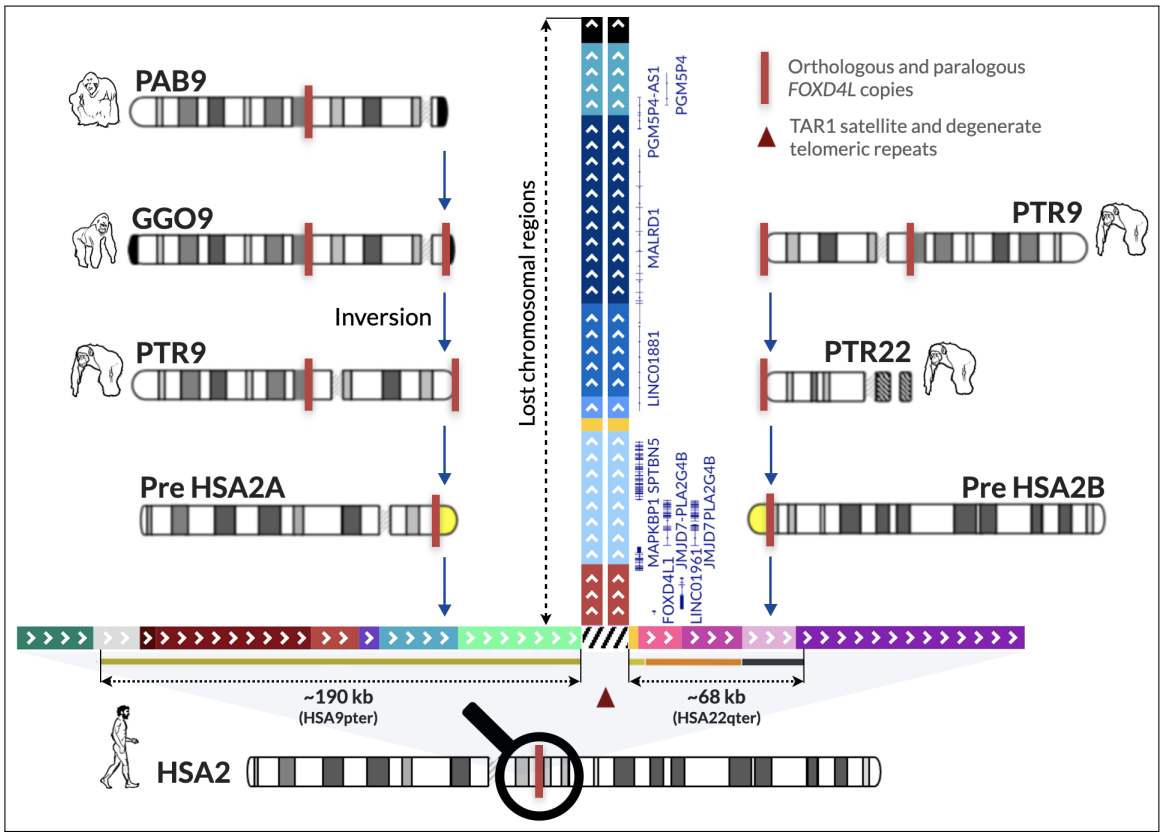
Jak opisano powyżej, redukcja liczby chromosomów z 48 w genomie małp człekokształtnych do 46 w genomie współczesnego człowieka jest z bardzo dużym prawdopodobieństwem wynikiem fuzji dwóch chromosomów mających swoje odpowiedniki u naczelnych innych niż człowiek. Sygnaturami tego zdarzenia, które można znaleźć na współczesnym genomie człowieka na chromosomie 2 jest obecność odwróconych powtórzeń telomerowych w miejscu fuzji oraz bloku zdegenerowanych sekwencji satelitarnych, który stanowi pozostałość centromeru. Dotychczasowe estymacje dotyczące czasu fuzji podają 4,5 miliona lat jako górną granicę czasu jej powstania.

Jedna z metod określenia czasu tej fuzji, zaproponowana w pracy Dreszer *et al.* (2007), polega na kwantyfikacji zdarzeń związanych ze zjawiskiem nazywanym w literaturze naukowej jako Biased Gene Conversion (BGC). Zjawisko to występuje podczas rekombinacji (Strathern *et al.*, 1995) i jest konsekwencją częstszego używania par nukleotydów silnie



Rysunek 3: Architektura genomu w pobliżu miejsca fuzji na chromosomie 2 człowieka oraz regiony syntenicznе w genomach małp człekokształtnych i człowieka. Rysunek przedstawia kolejno: sekwencje chromosomów 2Apter i 2Bpter orangutana (PAB) i goryla (GGO); chromosomy 2Apter, 2Bpter, 9pter, 12pter i 22qter szympansa (PTR) i bonobo (PPA); chromosom drugi człowieka z odpowiadającymi mu anotacjami sekwencji kodujących. Na każdym z zaprezentowanych chromosomów, unikalne kolory zostały użyte do oznaczenia homologicznych/ortologicznych fragmentów genetycznych (ang. contig), zgodnych także pomiędzy różnymi chromosomami i gatunkami. Współrzędne oznaczone są przy użyciu złożenia *hg38* genomu człowieka, a strzałki wskazują kierunek nici DNA. Ciemnoszare fragmenty z białymi krzyżykami reprezentują silnie mozaikowate segmentalne duplikacje lub powtórzenia tandemowe, których nie można przedstawić graficznie w czytelny sposób. Brązowe groty strzałek oznaczają sekwencję satelitarną TAR1 i powtórzenia telomerowe w miejscu fuzji HSA2 oraz ich ortologi w genomie małp człekokształtnych. Poniżej każdej sekwencji chromosomowej różnokolorowe paski reprezentują: (i) zielony - fragmenty genomu zrekonstruowane przy pomocy asemblera PhaseDancer nieobecne w najnowszych genomach referencyjnych, (ii) różowy - regiony o wysokiej homologii między chromosomami 2Apter i 2Bpter, prawdopodobnie mediujące zdarzenie fuzji, (iii) szary - region, który został utracony po fuzji w odniesieniu do HSA2. Na dole figury wskazano segmentalne duplikacje z okolic regionu fuzji, w tym fragment o długości ∼190 kb homologiczny do HSA9pter i trzy fragmenty o rozmiarze ∼ 68 kb w homologiczne do chromosomu HSA22pter. Fragment w kolorze lazurowym (chr2:113.523-113.554 kb) został zamplifikowany ∼ 400 razy w genomie szympansa (Cheng *et al.*, 2005).

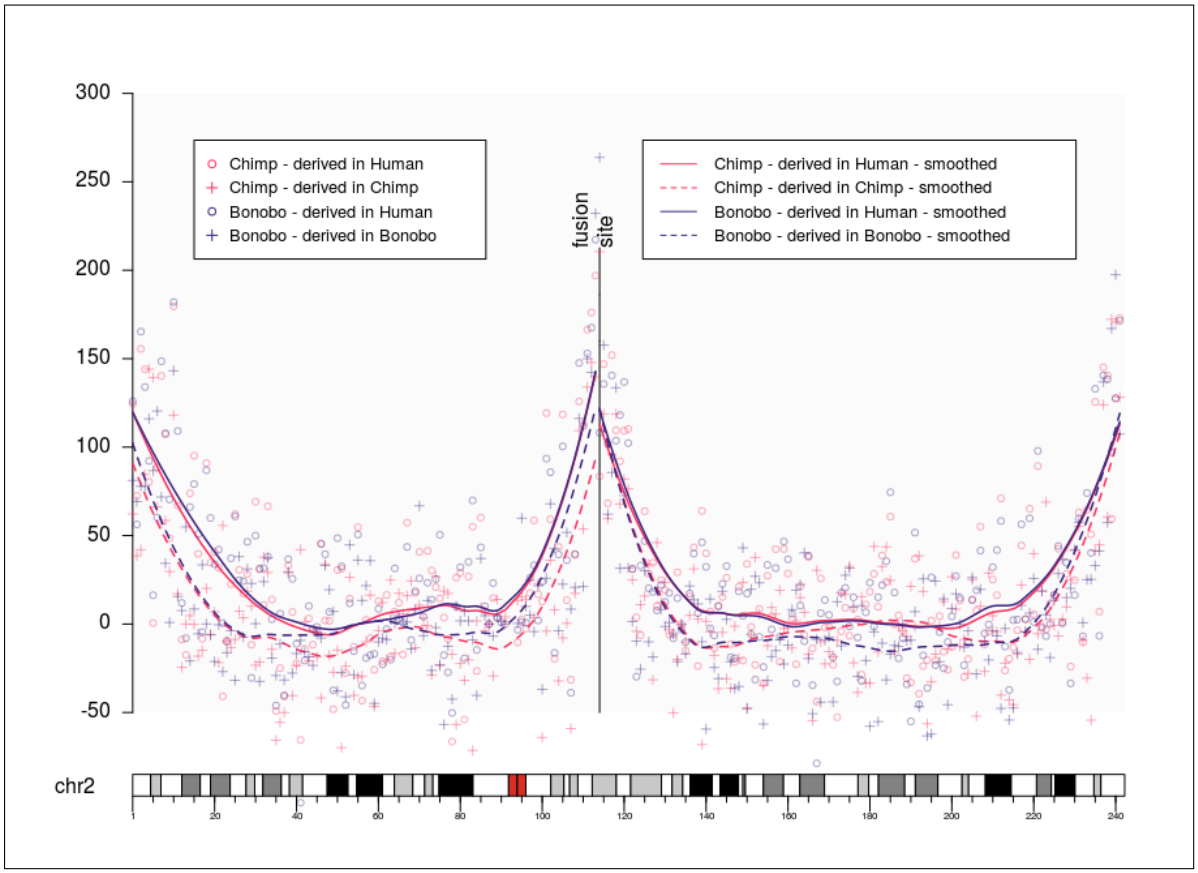
stabilnych (G,C) niż słabo stabilnych (A, T) w miejscach heterozygorycznych w heterodupleksowym DNA w procesie jego naprawy (Meunier and Duret, 2004). W publikacji Dreszer *et al.* (2007) zauważony został fakt, że BGC jest lokalnie nadreprezentowane w pobliżu



Rysunek 4: Model zdarzenia fuzji prowadzącej do powstania drugiego chromosomu człowieka zaproponowany na podstawie odtworzenia fragmentów subtelomerowych wzbogaconych w segmentalne duplikacje w genomach małp człekokształtnych, a nieobecnych w genomach referencyjnych tych gatunków. Miejsce fuzji jest otoczone z lewej strony przez segmentalną duplikację długości ~190 kb homologiczną do chromosomu 9p24.3, a z prawej strony przez segmentalną duplikację o długości ~68 kb homologiczną chromosomu 22q13.33 człowieka (z odpowiednio 98,9% i 97,8-99,1% identycznością sekwencji). Fragment długości ~190 kb zawierający gen *FOXD4L1* (oznaczony jako czerwony prostokąt), przypuszczalnie pochodzący z fragmentu syntenicznego do chromosomu 9q21.11 w genomie człowieka, uległ, jak wcześniej wykazano, transpozycji duplikującej na chromosome PTR2Apter po rozdzieleniu goryla od wspólnego przodka szympansa i człowieka (Martin *et al.*, 2002; Ventura *et al.*, 2012; Lese *et al.*, 1999; Wong *et al.*, 2004). Obie kopie flankują pericentromeryczną inwersję w genomach człowieka i szympansa, która powstała po oddzieleniu się do wspólnego przodka z gorylem (Martin *et al.*, 2002; Fan *et al.*, 2002; Wong *et al.*, 2004). W zaproponowanym modelu, fragment kopii PTR9pter został również skopiowany na chromosom PTR22qter, a później na PTR2Bter przed rozdzieleniem od wspólnego wspólnego przodka z gorylem i szympansem (Martin *et al.*, 2002; Fan *et al.*, 2002; Ning *et al.*, 1996; Wong *et al.*, 1999). Asemblacja regionów subtelomerowych wykazała istnienie długich fragmentów homologicznych (~190 kb) zawierających utracone fragmenty chromosomów praczłowieka przed fuzją 2Apter (pre-HSA2A) i 2Bpter (pre-HSA2B), które mogły posłużyć jako substrat mediujący translokacje podczas mejozy. Fuzja nastąpiła łącząc fragmenty zakończone sekwencją satelitarną TAR1 i powtórzeniami telomerowymi obecnymi zarówno w chromosomach pre-HSA2Apter, jak i pre-HSA2Bpter. Submikroskopowe subtelomeryczne rearanżacje u ludzi są stosunkowo częstą przyczyną chorób genetycznych powodujących opóźnienie w rozwoju lub niepełnosprawność intelektualną (Flint *et al.*, 1995). Analiza sekwencji wykazała, że dwie kopie sześciu genów kodujących białka (*FOXD4L1*, *JMJD7-PLA2G4B*, *MAPKBP1*, *PGM5P4*, *SPTBN5*, *CBWD2* i *MALRD1*) oraz trzech lncRNA (*LINC01881* and *LINC01961*, and *PGM5P4-AS1*) zostały utracone w czasie fuzji.

telomerów chromosomów autosomalnych. Korzystając ze statystyki Unexpected Bias Clustered Substitutions (UBCS) mierzącej wzbogacenie substytucji silnie stabilnych w stosunku do słabo stabilnych wśród substytucji sklastrowanych i porównując ich redukcję w regionach w pobliżu miejsca fuzji z ortologicznymi miejscami subtelomerowymi chromosomów 2a i 2b szympansa, autorzy oszacowali czas fuzji na 0,74 milionów lat temu z 95% przedziałem ufności 0–2,81.

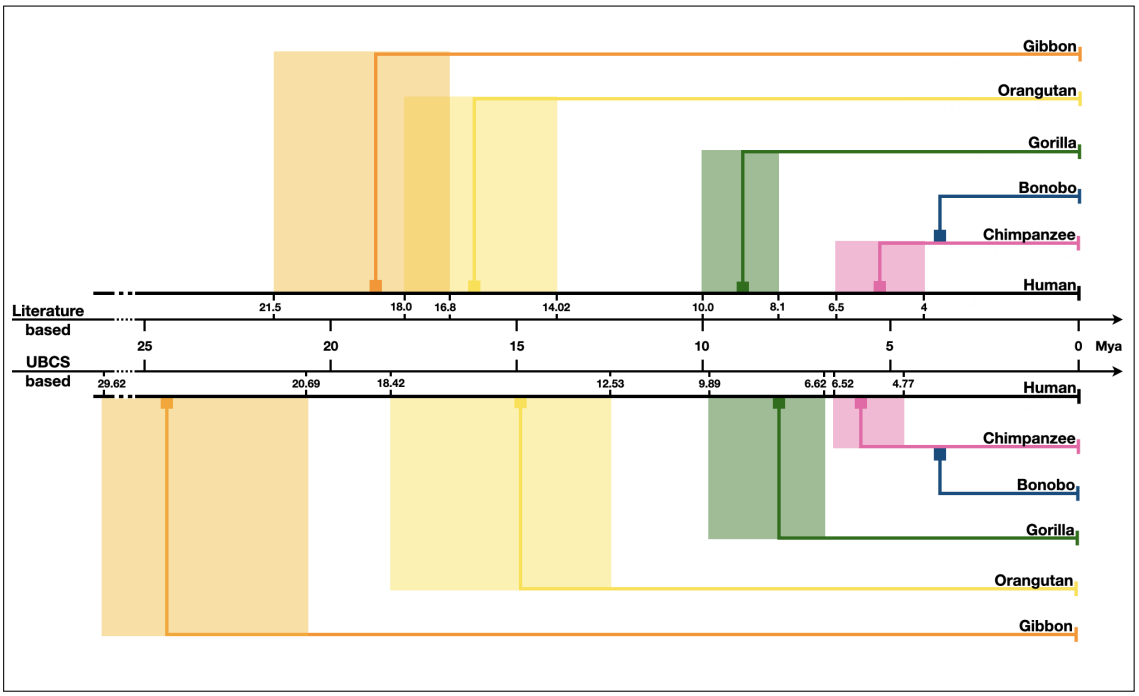
Niemniej jednak procedura obliczania wartości UBCS zaproponowana przez Dreszer



Rysunek 5: Wartości statystyki UBCS dla chromosomu 2 człowieka. Powyższa figura prezentuje wartości statystyki UBCS dla chromosomu 2 człowieka. Pionowa kreska oznacza punkt miejsca fuzji (chr2:113,600,000). Kolor linii odpowiada gatunkowi (szympansowi lub bonobo), dla którego obliczana jest wartość UBCS . Wartości linii ciągłej oznaczają wartości UBCS wyprowadzone dla człowieka, natomiast przerywanej dla szympansa lub bonobo.

et al. (2007) jest ściśle ograniczona i uwzględnia okna (fragmenty genomu) o wielkości 300 par zasad rozpoczynające się co 150 par zasad. To uproszczenie mogło prowadzić do niewłaściwych wyników, zwłaszcza w rejonach subtelomerowych zawierających izochory wzbogacone w nukleotydy G i C (Costantini *et al.*, 2006). Aby rozwiązać ten problem, opracowany został udoskonalony algorytm pozwalający na ponowne przeliczenie statystyki UBCS i podanie jej dokładnej wartości dla każdego możliwego przesunięcia okna. Zmieniony algorytm iteruje substytucje sumując ich udział w wartości UBCS dla danego regionu. Dla każdej substytucji wszystkie okna, które ją zawierają są kompresowane do równoważnego im wektora liczb naturalnych. Następnie, zastosowane technik programowania dynamicznego wyprowadzonych z zasady "włączeń i wyłączeń" oraz twierdzenia o prawdopodobieństwie całkowitym pozwoliło na określenie dokładnej wartości statystyki UBCS. Analizując ustalone na nowo wartości UBCS, określono, że fuzja powstała około 800,000 lat temu, z przedziałem ufności kończącym się na 2 milionach lat temu (patrz Rysunek 5).

Na podstawie statystyk pochodzących z ulepszonego algorytmu zaproponowana została metoda datowania odległości ewolucyjnej między dwoma gatunkami. W tym celu wprowadzono następującą miarę proporcji UBCS, zakładającą, że dane są dwie sekwencje genomu



Rysunek 6: Odległości ewolucyjne między genomami małp człekokształtnych a człowieka. Na figurze przedstawiono graficznie możliwe czasy specjacji oparte o najnowsze doniesienia literaturowe. Dla każdego gatunku minimalne i maksymalne datowanie jest zaznaczone na poziomej osi czasu. Korzystając z proporcji statystyk UBCS, oszacowane zostały czasy dla zdarzeń oddzielenia się od najniższego wspólnego przodka pomiędzy człowiekiem a małpami człekokształtnymi. Wynoszą one odpowiednio, dla szympansa: 4,77-6,52 miliona lat, dla bonobo: 4,35-5,85 miliona lat, dla goryla: 6,62-9,89 miliona lat, dla orangutana: 12,53-18,42 miliona lat, dla gibona: 20,68-29,62 miliona lat. Należy zauważyć, że wszystkie te wartości pokrywają się z przedziałami, których granice są określone przez odpowiednie datowania uzyskane przy pomocy innych metod zaczerpniętych z literatury.

$\mathcal{G}_x$ i $\mathcal{G}_y$, N chromosomów oraz M okien o rozmiarze 1 Mb w regionach telomerowych każdego chromosomu. Oznaczamy $\mathcal{G}_{x_j^i}$ jako j-te okno na i-tym chromosomie genomu $\mathcal{G}_x$, a wartość jego statystyki UBCS jako $\mathcal{U}(\mathcal{G}_{x_j^i})$ oraz $\overline{x}$ jako odwróconą komplementarną sekwencję x (pierwsze okno $\overline{x}$ to ostatni 1 Mb x). Średnią proporcję UBCS między telomerami na ramionach p i q i-tego chromosomu genomów $\mathcal{G}_x$ i $\mathcal{G}_y$ wyraża się wzorem:

$$\mathcal{T}_p(i) = \frac{\sum_{j=1}^{M} \mathcal{U}(\mathcal{G}_{x_j^i})}{\sum_{j=1}^{M} \mathcal{U}(\mathcal{G}_{y_j^i})} \qquad \mathcal{T}_q(i) = \frac{\sum_{j=1}^{M} \mathcal{U}(\mathcal{G}_{\overline{x}_j^i})}{\sum_{j=1}^{M} \mathcal{U}(\mathcal{G}_{\overline{y}_j^i})}$$

a odległość ewolucyjną szacowana jest na podstawie średniej proporcji UBCS między genomami $\mathcal{G}_x$ and $\mathcal{G}_y$ jako:

$$\mathcal{G}_x||\mathcal{G}_y = \text{median}(\{\mathcal{T}_p(i) : i \in \mathcal{CT}_p\} \cup \{\mathcal{T}_q(i) : i \in \mathcal{CT}_q\})$$

gdzie $\mathcal{CT}_p$ i $\mathcal{CT}_q$ są tak zwanymi *chromosomami kontrolnymi* użytymi do określania proporcji wartości UBCS pomiędzy ramionami odpowiednio p and q.

Przedział ufności dla proporcji UBCS jest określany przy pomocy metody *bootstrap* poprzez losowanie ze zwracaniem podzbioru telomerów i podzbioru okien długości 1 Mb. Procedura jest powtórzana 1000 razy, a następnie eliminowane jest 5% najbardziej odstających wartości. Data specjacji jest ustalana przez pomnożenie proporcji UBCS, która kwan-

tyfikuje stosunek odległości ewolucyjnej pomiędzy pomiędzy dwoma gatunkami o genomach $\mathcal{G}_x$ i $\mathcal{G}_y$, poprzez znaną datę specjacji trzeciego gatunku i gatunku o genomie $\mathcal{G}_y$.

Za pomocą tej metody zrekonstruowane zostały odległości ewolucyjne między małpami człekokształtnymi (*Hominoide*). Czasy specjacji szympansa i bonobo oszacowano w niewielkiej odległości czasowej, odpowiednio między 4,7-6,6 milionów lat temu i 5,5-7,5 milionów lat temu. W przypadku goryla, orangutana i gibona oszacowanie te wynoszą odpowiednio 6,6-9,9 mln lat, 12,5-18,4 mln lat, 20,7-29,6 mln lat (patrz Rysunek 6). Warto zauważyć, że wszystkie te wartości pokrywają się z przedziałami, których granice są określone przez datowania specjacji uzyskane przy pomocy innych metod zaczerpniętych z literatury (Chan *et al.*, 2010; Carbone *et al.*, 2014; Chatterjee *et al.*, 2009; Gronau *et al.*, 2011; Scally *et al.*, 2012; Stone *et al.*, 2010).

Podsumowując, rezultaty przedstawione w rozdziale trzecim rozprawy doktorskiej rzucają nowe światło na określenie czasu fuzji chromosomów prowadzącej do powstania chromosomu 2 człowieka oraz dostarczają nowej alternatywy obliczeniowej do datowania zdarzeń specjacji. Rezultaty te zostały zaprezentowane na zdalnej konferencji *16th International Symposium on Bioinformatics Research and Applications (ISBRA)* i opublikowane w czasopiśmie *BMC genomics* (Poszewiecka *et al.*, 2022a).

Wydajny algorytm enumeracji minimalnych liniowych eulerowskich dekompozycji grafu kariotypowego

Ewolucja jest ciągłym procesem zachodzącym również obecnie, a jednym z jej mechanizmów są rearanżacje chromosomowe najczęściej występujące w formie duplikacji, delecji, inwersji i translokacji. Złożone rearanżacje chromosomowe to zmiany strukturalne obejmujące więcej niż dwa punkty złamania. Powodują one zmianę kolejności, orientacji lub liczby kopii fragmentów chromosomów ograniczonych przez punkty złamania. Gdy zmiany te amplifikują materiał genetyczny lub dotyczą chromosomów homologicznych ułożenie fragmentów w chromosomach pochodnych może nie być jednoznacznie scharakteryzowane wyłącznie na podstawie punktów złamań rearanżacji. Badane rearanżacje chromosomowe najczęściej powstają *de novo* i z tego względu są pozbawione polimorfizmów pojedynczego nukleotydu oraz małych insercji lub delecji. Uniemożliwia to rozróżnienie od siebie kopii fragmentów pochodzących z rożnych części chromosomów pochodnych. Istotna staje się więc potrzeba zaprojektowania wydajnego algorytmu, który wylicza wszystkie możliwe scenariusze złożonych rearanżacji chromosomowych na podstawie punktów ich załamań. Lista możliwych scenariuszy złożonych rearanżacji może zostać poddana dalszym analizom wyjaśniającym ich konsekwencje molekularne.

W tym celu zaprezentowany został wydajny algorytm listujący wszystkie możliwe scenariusze złożonych rearanżacji chromosomowych. Takie rearanżacje są reprezentowane

przez tak zwany graf kariotypowy użyty w pracy Aganezov *et al.* (2019). Wierzchołki w tym grafie odpowiadają początkowi bądź końcowi segmentu genomu, a krawędzie są dwóch typów: segmentalne i sąsiedztwa. Pierwsze z nich kodują segmenty (rozłączne fragmenty genomu), natomiast drugie przejścia pomiędzy segmentami. Graf kariotypowy to z definicji multigraf, a liczbę kopii krawędzi w takim grafie określamy jako jej krotność. Kolekcje ścieżek i cykli składających się z naprzemiennie ułożonych krawędzi segmentalnych i sąsiedztwa, gdzie liczba wystąpień każdej krawędzi jest równa jej krotności nazywana jest eulerowską dekompozycją grafu kariotypowego. Powszechnie znanym jest fakt, że minimalna pod względem liczności eulerowska dekompozycja grafu kariotypowego złożona jest wyłącznie z liniowych chromosomów, jeżeli nie zawiera spójnych składowych bez telomerów. Każda eulerowska dekompozycja grafu kariotypowego określa pewien scenariusz rearanżacji chromosomowej.

W rozdziale czwartym rozprawy doktorskiej przedstawiony został algorytm wylistowania wszystkich Minimalnych Uporządkowanych Eulerowskich Dekompozycji Liniowodekomponowalnego Grafu Kariotypowego. W tym celu przeformułowano problem do równoważnego problemu wylistowania ścieżek Eulera złożonych z krawędzi o naprzemiennych typach spełniających pewne dodatkowe własności w Rozszerzonym Liniowodekomponowalnym Grafie Kariotypowym (ALDKG) zbudowanym na podstawie wejściowego grafu. Rozwiązanie tego problemu można zdefiniować rekurencyjnie jako przedłużenie ustalonej ścieżki o krawędzie, dla których zmniejszenie krotności o jeden nie prowadzi do powstania nietrywialnych spójnych składowych zawierających elementy bez telomerów w wynikowym grafie. Aby zapewnić, że rekurencja będzie wydajna. zdefiniowane zostały dwie struktury danych: *certyfikat spójności* i *certyfikat świadka. Certyfikat spójności* pozwala odpowiadać na pytania, czy zmniejszenie krotności danej krawędzi rozspójnia graf w niepożądany sposób. Własności *certyfikatu świadka* pozwoliły z kolei udowodnić, że tylko jeden z wierzchołków incydentnych z przetwarzanym wierzchołkiem może rozspójnić graf w taki sposób. Dzięki temu liczba zapytań do *certyfikatu spójności* podczas generowania dekompozycji mogła być istotnie ograniczona, co wpłynęło na zmniejszenie złożoności czasowej algorytmu. Zastosowanie *certyfikatu spójności* i wykorzystanie pewnych własności grafów kariotypowych pozwoliło na przechodzenie przez drzewo rekurencji w sposób unikający przeglądania wierzchołków, które nie prowadzą do wygenerowania jakiejś dekompozycji. Wykorzystując powyższe koncepcje, zaprojektowany został algorytm, który wylicza wszystkie minimalne liniowe dekompozycje eulerowskie grafów kariotypowych z wielomianową złożonością opóźnienia czasowego $O(log(n)^2 \cdot l)$, gdzie n jest liczbą wierzchołków w grafie kariotypowym, a l jest długością dekompozycji grafu.

W analizie złożoności czasowej i pamięciowej algorytmu kluczowe było następujące twierdzenie, którego dowód znajduje się w rozprawie doktorskiej:

Twierdzenie. *Mając dany Rozszerzony Liniowodekomponowalny Graf Kariotypowy (ALDKG) wszystkie jego Minimalne Uporządkowane Eulerowskie Dekompozycje odpowiadające Minimalnym Uporządkowanym Eulerowskim Dekompozycjom wejściowego Liniowodekomponowalnego Grafu Kariotypowego (LDKG) mogą być wylistowane w porządku leksykograficznym (gdzie alfabetem jest zbiór krawędziowych reprezentacji ścieżek) bez duplikatów w złożoności czasowej $O(log^2(n) \cdot l)$ dla pojedynczej dekompozycji. Inicjalizacja struktur danych ma złożoność czasową $O(log^2(n) \cdot m)$, a pamięciowa złożoność algorytmu to $O(log(m) \cdot l)$, gdzie n jest liczbą wierzchołków w grafie kariotypowym, m liczbą krawędzi, a l jest długością dekompozycji grafu.*

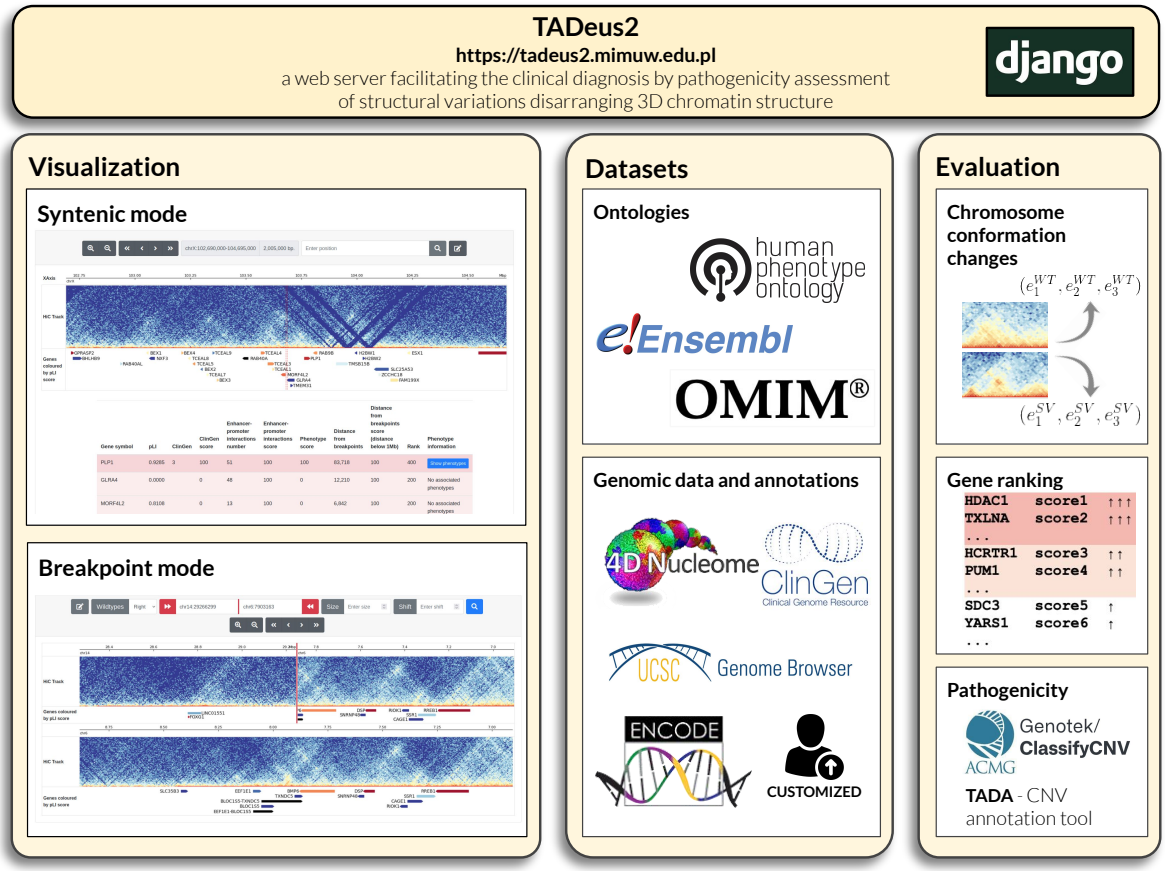
Wykazana została również użyteczność tego algorytmu do przedstawienia prawdopodobnych scenariuszy złożonej rearanżacji chromosomowej u pacjenta, którego przypadek został przedstawiony w studium Nazaryan-Petersen *et al.* (2018).

Interpretacja zmian strukturalnych zaburzających trójwymiarową strukturę chromatyny

Jak wspomniano we wstępie, na przestrzeni ostatnich lat dokonał się ogromny postęp w identyfikacji strukturalnych wariantów w genomie człowieka. Jednak ich kliniczna interpretacja, zwłaszcza gdy występują w niekodującym DNA, pozostaje wyzwaniem. Jednym z powodów jest brak odpowiednich narzędzi uwzględniających zmiany w trójwymiarowej architekturze chromatyny powodowanych przez warianty strukturalne.

By wypełnić tę lukę, w ramach rozprawy doktorskiej, stworzona została aplikacja internetowa `TADeus2` dedykowana szybkiej ocenie wpływu zmian w konformacji chromatyny poprzez wizualizację strukturalnych wariantów obejmujących domeny topologiczne (patrz Rysunek 7). `TADeus2` udostępnia innowacyjną przegądarkę całogenomową umożliwiającą wygodną wizualizację strukturalnych wariantów, zarówno w zwykłym trybie, jak i z perspektywy punktów złamań rearanżacji. Tryb ten, nazywany *breakpoint view*, umożliwia przedstawianie dwóch fragmentów genomu połączonych punktem złamania rearanżacji, jednocześnie pozwalając na podgląd anotacji fragmentów genomu referencyjnego po obydwu stronach punktu złamania Należy podkreślić, że jest to pierwsza przeglądarka genomowa z tego rodzaju funkcjonalnością.

Co więcej, aplikacja `TADeus2` powala na kwantyfikację i dostarcza ranking patogeniczności strukturalnych wariantów przy użyciu narzędzi TADa (Hertzberg *et al.*, 2022) i ClassifyCNV (Gurbich and Ilinsky, 2020). Warto nadmienić, że drugie narzędzie oblicza wskaźnik patogeniczności dla polimorfizmów liczby kopii zgodnie z wytycznymi American College of Medical Genetics. Ponadto zaproponowano oryginalną, opartą na próbkowaniu metodę obliczania p-wartości, określającą ilościowo patogeniczność wariantu na podstawie

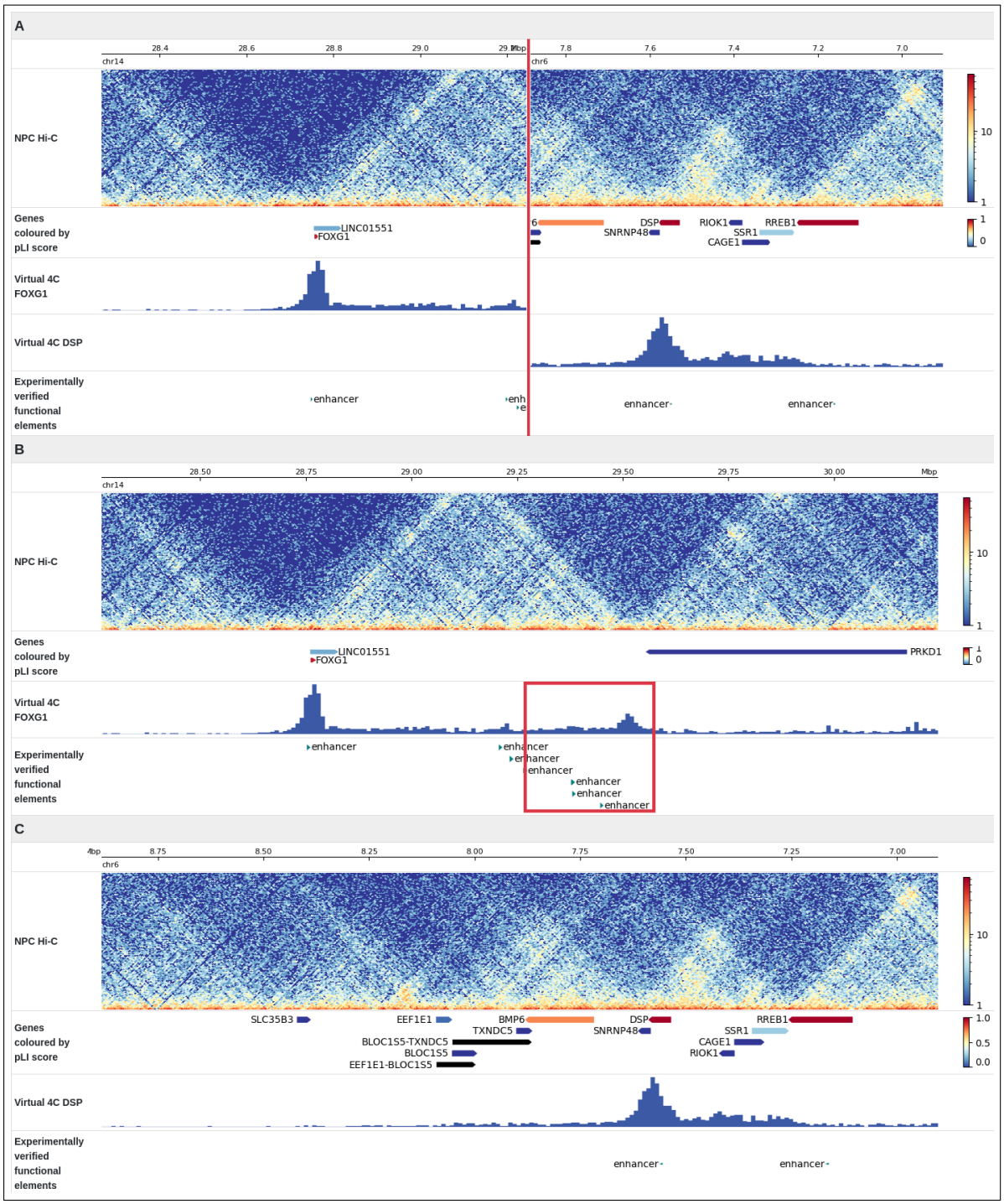


Rysunek 7: Główne funkcjonalności oferowane przez aplikację internetową `TADeus2`. Główne funkcjonalności oferowane przez aplikację internetową `TADeus2` wraz ze sposobem ich wykorzystania w klinicznej ocenie patogeniczności zmian strukturalnych zaburzających trójwymiarową strukturę chromatyny.

liczby przerwanych przewidywanych interakcji enhancera z promotorem.

Ponadto zaproponowano ranking genów w okolicy strukturalnych wariantów ze względu na ich przewidywaną patogeniczność. Używa on następujących charakterystyk dla genów w pobliżu punktów złamań: wskaźnika ClinGen (Rehm *et al.*, 2015) wrażliwości na dawkę (haploinsufficiency/triplosensitivity), liczbę zakłóconych przewidywanych interakcji łączących potencjalny enhancer z promotorem genu, liczbę związanych z genem wpisów w bazie danych Human Phenotype Ontology (Köhler *et al.*, 2016), oraz jego odległość od punktów złamań rearanżacji. Ranking został zwalidowany przy użyciu 21 opisanych w literaturze przypadków efektu pozycji spowodowanych przez warianty strukturalne i polimorfizmy liczby kopii. We wszystkich przypadkach geny przyczyniające się do choroby zostały sklasyfikowane jako mocni (18; 85,7%) lub prawdopodobni (3; 14,3%) kandydaci do spowodowania efektu pozycji.

Zaproponowany sposób oceny patogeniczności wariantów strukturalnych został z sukcesem użyty do analizy efektu pozycji u czterech przypadków pacjentów cierpiących na różne schorzenia genetyczne. W każdym z tych przypadków zaproponowano molekularne przyczyny zespołów genetycznych. Innowacyjny tryb wizualizacji punktów złamań rearanżacji



Rysunek 8: Analiza zrównoważonej translokacji 46,XX,t(6;14)(p25.1;q12) u pacjenta z napadami padaczkowymi i znacznym opóźnieniem rozwojowym przy pomocy przeglądarki genomowej w trybie *breakpoint view*. Wizualizacja punktu złamania na chromosomie pochodnym po translokacji (**A**W analizie złożoności czasoW analizie złożoności czaso) i chromosomy referencyjne: chromosom 14 (**B**) oraz chromosom 6 (**C**). Do wygenerowania ścieżki przedstawiającej architekturę chromatyny wykorzystane zostały dane NPC Hi-C (górna ścieżka). Ponadto rysunek zawiera następujące ścieżki: wskaźnik pLI gnomAD genów (ścieżka dolna) w skali kolorów (0 = niebieski i 1 = czerwony), dane Hi-C wyświetlone w trybie *virtual 4C* dla regionów promotora (+/- 2,5 kb od TSS) dla *FOXG1* i *DSP* oraz zweryfikowane eksperymentalnie elementy funkcjonalne. Enhancery przemieszczone przez translokację zaznaczono czerwonym prostokątem.

został użyty między innymi do analizy efektu pozycji wywołanego przez zbalansowaną translokację (46,XX,t(6;14)(p25.1;q12)) w sąsiedztwie genu *FOXG1* u pacjenta z napadami

padaczkowymi i ciężkim opóźnieniem rozwoju. Użycie przeglądarki w trybie *breakpoint view* pozwoliło na wizualizację fragmentów genomu sąsiadujących z punktem złamania i wskazanie dwóch enhancerów, których przemieszczenie spowodowało najprawdopodobniej chorobowy fenotyp (patrz Rysunek 8). Ponadto ranking genów zastosowany w aplikacji `TADeus2` poprawnie wskazał gen, którego zmieniona ekspresja odpowiedzialna była za chorobę jako mocnego kandydata do spowodowania efektu pozycji.

Pierwszą wersję aplikacji internetowej zaprezentowano na konferencji *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Została ona również opublikowana w materiałach pokonferencyjnych (Poszewiecka *et al.*, 2018). Druga, rozszerzona wersja aplikacji, została opisana w czasopiśmie *Nucleic Acids Research* (Poszewiecka *et al.* (2022b)) i jest publicznie dostępna pod adresem https:/ /tadeus2.mimuw.edu.pl. `TADeus2` and i jego wcześniejsza wersja `TADeus` zostały użyte w niedawno opublikowanych artykułach naukowych (Pienkowski *et al.*, 2019, 2020).

Wykaz publikacji głównych wyników przedstawionych w rozprawie

Poszewiecka, B., Gogolewski, K., Karolak, J.A., Stankiewicz, P. and Gambin, A., 2023. From 48 to 46 chromosomes: a novel targeted assembler of segmental duplications unravels the complexity of the HSA2 fusion. *Genome Biology*, accepted for publication by the editor.

Poszewiecka, B., Pienkowski, V.M., Nowosad, K., Robin, J.D., Gogolewski, K. and Gambin, A., 2022. TADeus2: a web server facilitating the clinical diagnosis by pathogenicity assessment of structural variations disarranging 3D chromatin structure. *Nucleic Acids Research*, 50(W1), pp. W744–W752,

Poszewiecka, B., Gogolewski, K., Stankiewicz, P. and Gambin, A., 2022. Revised time estimation of the ancestral human chromosome 2 fusion. *BMC genomics*, 23(6), pp.1-16.

Poszewiecka, B., Stankiewicz, P., Gambin, T. and Gambin, A., 2018, December. TADeus-a tool for clinical interpretation of structural variants modifying chromatin organization. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 84-87). IEEE.

Lista innych publikacji

Pienkowski, V.M., Kucharczyk, M., Młynek, M., Szczałuba, K., Rydzanicz, M., Poszewiecka, B., Skórka, A., Sykulski, M., Biernacka, A., Koppolu, A.A. and Posmyk, R., 2019. Mapping of breakpoints in balanced chromosomal translocations by shallow whole-genome sequencing points to EFNA5, BAHD1 and PPP2R5E as novel candidates for genes causing human Mendelian disorders. *Journal of Medical Genetics*, 56(2), pp.104-112.

Murcia Pienkowski, V., Kucharczyk, M., Rydzanicz, M., Poszewiecka, B., Pachota, K., Młynek, M., Stawiński, P., Pollak, A., Kosińska, J., Wojciechowska, K. and Lejman, M., 2020. Breakpoint mapping of symptomatic balanced translocations links the EPHA6, KLF13 and UBR3 genes to novel disease phenotype. *Journal of Clinical Medicine*, 9(5), p.1245.

Bibliografia

Aganezov, S., Zban, I., Aksenov, V., Alexeev, N. and Schatz, M. C. (2019). Recovering rearranged cancer chromosomes from karyotype graphs. *BMC bioinformatics*, **20** (20), 1–11.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, **21** (1).

Carbone, L., Harris, R. A., Gnerre, S., Veeramah, K. R., Lorente-Galdos, B., Huddleston, J., Meyer, T. J., Herrero, J., Roos, C., Aken, B., Anaclerio, F., Archidiacono, N., Baker, C., Barrell, D., Batzer, M. A., Beal, K., Blancher, A., Bohrson, C. L., Brameier, M., Campbell, M. S., Capozzi, O., Casola, C., Chiatante, G., Cree, A., Damert, A., de Jong, P. J., Dumas, L., Fernandez-Callejo, M., Flicek, P., Fuchs, N. V., Gut, I., Gut, M., Hahn, M. W., Hernandez-Rodriguez, J., Hillier, L. W., Hubley, R., Ianc, B., Izsv?k, Z., Jablonski, N. G., Johnstone, L. M., Karimpour-Fard, A., Konkel, M. K., Kostka, D., Lazar, N. H., Lee, S. L., Lewis, L. R., Liu, Y., Locke, D. P., Mallick, S., Mendez, F. L., Muffato, M., Nazareth, L. V., Nevonen, K. A., O'Bleness, M., Ochis, C., Odom, D. T., Pollard, K. S., Quilez, J., Reich, D., Rocchi, M., Schumann, G. G., Searle, S., Sikela, J. M., Skollar, G., Smit, A., Sonmez, K., ten Hallers, B., Terhune, E., Thomas, G. W., Ullmer, B., Ventura, M., Walker, J. A., Wall, J. D., Walter, L., Ward, M. C., Wheelan, S. J., Whelan, C. W., White, S., Wilhelm, L. J., Woerner, A. E., Yandell, M., Zhu, B., Hammer, M. F., Marques-Bonet, T., Eichler, E. E., Fulton, L., Fronick, C., Muzny, D. M., Warren, W. C., Worley, K. C., Rogers, J., Wilson, R. K. and Gibbs, R. A. (2014). Gibbon genome and the fast karyotype evolution of small apes. *Nature*, **513** (7517), 195–201.

Chan, Y. C., Roos, C., Inoue-Murayama, M., Inoue, E., Shih, C. C., Pei, K. J. and Vigilant, L. (2010). Mitochondrial genome sequences effectively reveal the phylogeny of Hylobates gibbons. *PLoS ONE*, **5** (12), e14419.

Chatterjee, H. J., Ho, S. Y., Barnes, I. and Groves, C. (2009). Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evol. Biol.*, **9**, 259.

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R. K., Pääbo, S., Rocchi, M. and Eichler, E. E. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, **437** (7055), 88–93.

Costantini, M., Clay, O., Auletta, F. and Bernardi, G. (2006). An isochore map of human chromosomes. *Genome Res*, **16** (4), 536–541.

Dreszer, T. R., Wall, G. D., Haussler, D. and Pollard, K. S. (2007). Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome research*, **17** (10), 1420–1430.

Fan, Y., Linardopoulou, E., Friedman, C., Williams, E. and Trask, B. J. (2002). Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14. 1 and paralogous regions on other human chromosomes. *Genome research*, **12** (11), 1651–1662.

Flint, J., Wilkie, A. O., Buckle, V. J., Winter, R. M., Holland, A. J. and McDermid, H. E. (1995). The detection of subtelomeric chromosomal rearrangements in idiopathic mental retardation. *Nature genetics*, **9** (2), 132–140.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.*, **43** (10), 1031–1034.

Gurbich, T. A. and Ilinsky, V. V. (2020). Classifycnv: a tool for clinical annotation of copy-number variants. *Scientific reports*, **10** (1), 1–7.

Hertzberg, J., Mundlos, S., Vingron, M. and Gallone, G. (2022). TADA—a machine learning tool for functional annotation-based prioritisation of pathogenic CNVs. *Genome Biology*, **23** (1).

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J. and Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, **24** (4), 688–696.

il Sohn, J. and Nam, J.-W. (2016). The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics*, p. bbw096.

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M. *et al.* (2016). The human phenotype ontology in 2017. *Nucleic acids research*, **45** (D1), D865–D876.

Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, **37** (5), 540–546.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, **27** (5), 722–736.

Lese, C. M., Fantes, J. A., Riethman, H. C. and Ledbetter, D. H. (1999). Characterization of physical gap sizes at human telomeres. *Genome research*, **9** (9), 888–894.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, **32** (14), 2103–2110.

Martin, C. L., Wong, A., Gross, A., Chung, J., Fantes, J. A. and Ledbetter, D. H. (2002). The evolutionary origin of human subtelomeric homologies—or where the ends begin. *The American Journal of Human Genetics*, **70** (4), 972–984.

Meunier, J. and Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.*, **21** (6), 984–990.

Nazaryan-Petersen, L., Eisfeldt, J., Pettersson, M., Lundin, J., Nilsson, D., Wincent, J., Lieden, A., Lovmar, L., Ottosson, J., Gacic, J. *et al.* (2018). Replicative and non-replicative mechanisms in the formation of clustered cnvs are indicated by whole genome characterization. *PLoS genetics*, **14** (11), e1007780.

Ning, Y., Rosenberg, M., Ledbetter, D. H. and Biesecker, L. G. (1996). Isolation of the human chromosome 22q telomere and its application to detection of cryptic chromosomal abnormalities. *Human genetics*, **97** (6), 765–769.

Pienkowski, V. M., Kucharczyk, M., Młynek, M., Szczałuba, K., Rydzanicz, M., Poszewiecka, B., Skórka, A., Sykulski, M., Biernacka, A., Koppolu, A. A. *et al.* (2019). Mapping of breakpoints in balanced chromosomal translocations by shallow whole-genome sequencing points to efna5, bahd1 and ppp2r5e as novel candidates for genes causing human mendelian disorders. *Journal of Medical Genetics*, **56** (2), 104–112.

—, —, Rydzanicz, M., Poszewiecka, B., Pachota, K., Młynek, M., Stawiński, P., Pollak, A., Kosińska, J., Wojciechowska, K., Lejman, M., Cieślikowska, A., Wicher, D., Stembalska, A., Matuszewska, K., Materna-Kiryluk, A., Gambin, A., Chrzanowska, K.,

Krajewska-Walasek, M. and Płoski, R. (2020). Breakpoint mapping of symptomatic balanced translocations links the EPHA6, KLF13 and UBR3 genes to novel disease phenotype. *Journal of Clinical Medicine*, **9** (5), 1245.

Poszewiecka, B., Gogolewski, K., Stankiewicz, P. and Gambin, A. (2022a). Revised time estimation of the ancestral human chromosome 2 fusion. *BMC Genomics*, **23** (S6).

—, Pienkowski, V. M., Nowosad, K., Robin, J. D., Gogolewski, K. and Gambin, A. (2022b). Tadeus2: a web server facilitating the clinical diagnosis by pathogenicity assessment of structural variations disarranging 3d chromatin structure. *Nucleic Acids Research*.

—, Stankiewicz, P., Gambin, T. and Gambin, A. (2018). TADeus-a tool for clinical interpretation of structural variants modifying chromatin organization. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE.

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L. *et al.* (2015). Clingen—the clinical genome resource. *New England Journal of Medicine*, **372** (23), 2235–2242.

Ruan, J. and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature methods*, **17** (2), 155–158.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E. V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M., Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Stenson, P. D., Turner, D. J., Vigilant, L., Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E., Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O. A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C. and Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483** (7388), 169–175.

Stone, A. C., Battistuzzi, F. U., Kubatko, L. S., Perry, G. H., Trudeau, E., Lin, H. and Kumar, S. (2010). More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **365** (1556), 3277–3288.

Strathern, J. N., Shafer, B. K. and McGill, C. B. (1995). DNA synthesis errors associated with double-strand-break repair. *Genetics*, **140** (3), 965–972.

Ventura, M., Catacchio, C. R., Sajjadian, S., Vives, L., Sudmant, P. H., Marques-Bonet, T., Graves, T. A., Wilson, R. K. and Eichler, E. E. (2012). The evolution of african great ape subtelomeric heterochromatin and the fusion of human chromosome 2. *Genome Research*, **22** (6), 1036–1049.

Vollger, M. R., Dishuck, P. C., Sorensen, M., Welch, A. E., Dang, V., Dougherty, M. L., Graves-Lindsay, T. A., Wilson, R. K., Chaisson, M. J. and Eichler, E. E. (2019a). Long-read sequence and assembly of segmental duplications. *Nature methods*, **16** (1), 88–94.

—, Guitart, X., Dishuck, P. C., Mercuri, L., Harvey, W. T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K. M., Lewis, A. P., Hoekzema, K., Porubsky, D., Li, R., Nurk, S., Koren, S., Miga, K. H., Phillippy, A. M., Timp, W., Ventura, M. and Eichler, E. E. (2022). Segmental duplications and their variation in a complete human genome. *Science*, **376** (6588).

—, Logsdon, G. A., Audano, P. A., Sulovari, A., Porubsky, D., Peluso, P., Wenger, A. M., Concepcion, G. T., Kronenberg, Z. N., Munson, K. M., Baker, C., Sanders, A. D., Spierings, D. C., Lansdorp, P. M., Surti, U., Hunkapiller, M. W. and Eichler, E. E. (2019b). Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Annals of Human Genetics*, **84** (2), 125–140.

Wong, A., Vallender, E. J., Heretis, K., Ilkin, Y., Lahn, B. T., Martin, C. L. and Ledbetter, D. H. (2004). Diverse fates of paralogs following segmental duplication of telomeric genes. *Genomics*, **84** (2), 239–247.

Wong, A. C., Shkolny, D., Dorman, A., Willingham, D., Roe, B. A. and McDermid, H. E. (1999). Two novel human rab genes with near identical sequence each map to a telomere-associated region: the subtelomeric region of 22q13. 3 and the ancestral telomere band 2q13. *Genomics*, **59** (3), 326–334.