Rozwój narzędzi informatycznych do analizy danych genomowych uzyskanych z mikromacierzy i sekwenatorów następnej generacji

Autoreferat

Tomasz Gambin

21.12.2017

Spis treści

1. Dane osobowe i przebieg zatrudnienia w jednostkach naukowych

Imię i Nazwisko: Tomasz Gambin

Dyplomy i stopnie naukowe:

2012 Warszawa - Doktor nauk technicznych, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych. Tytuł rozprawy: Design of Experiments and Genomic Data Analysis in Array-based CGH Technology". (Projektowanie eksperymentów i analiza danych genomowych w technologii mikromacierzy CGH).

2007 Warszawa - Magister inżynier, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych

Przebieg zatrudnienia w jednostkach naukowych:

2012.03-obecnie: Politechnika Warszawska, Instytut Informatyki, adiunkt

2015.03-obecnie: Zakład Genetyki Medycznej, Instytut Matki i Dziecka w Warszawie, asystent-specjalista

2013.01-2014.12: Baylor College of Medicine, Houston, Texas, Stany Zjednoczone, staż podoktorski

2012.07-2012.10: Staż podoktorski realizowany dzięki stypendium w ramach projektu "Technologie informacyjne: badania i ich interdyscyplinarne zastosowania" (UDAPOKL.04.01.01-00-051/10-00); projekt współfinansowany ze środków Unii Europejskiej, realizowany w Instytucie Biochemii i Biocybernetyki Polskiej Akademii Nauk, Warszawa, Polska

2008.09-2008.11: Baylor College of Medicine, Houston, Texas, Stany Zjednoczone, staż doktorancki

2007.10-2012.03: Politechnika Warszawska, Instytut Informatyki, doktorant

2. Osiągnięcie naukowe

W skład osiągnięcia naukowego wchodzi cykl jedenastu publikacji **[P1-P11]** oraz pięć narzędzi informatycznych **[T1-T5]**.

2.1 Tytuł osiągnięcia naukowego

"Rozwój narzędzi informatycznych do analizy danych genomowych uzyskanych z mikromacierzy i sekwenatorów następnej generacji"

2.2 Wykaz publikacji stanowiących osiągnięcie naukowe i opis wkładu własnego

Pozycja nazwisk autorów odzwierciedla system stosowany w naukach biologicznych i medycynie. Pierwszy autor jest osobą odpowiedzialną za wykonanie największego wkładu pracy badawczej, ostatni (ang. senior) autor jest osobą, która nadzoruje badania oraz posiada decydujący wpływ przy definiowaniu celu i zakresu pracy. Przy równym wkładzie pracy pierwszych autorów, są oni oznaczeni (symbolem "*") jako "wspólny pierwszy autor". Większość współautorów prac wchodzących w skład osiągnięcia naukowego stanowili biolodzy molekularni (odpowiedzialni za wykonanie eksperymentów) oraz lekarze genetycy (odpowiedzialni za kontakt z pacjentami).

[P1] P Dittwald*, **T Gambin***, P Szafrański, J Li, S Amato, M Y Divon, L X Rodríguez Rojas, L E Elton, D A Scott, C P Schaaf, W Torres-Martinez, A K Stevens, J A Rosenfeld, S Agadi, D Francis, S-H L Kang, A Breman, S R Lalani, C A Bacino, W Bi, A Milosavljevic, A L Beaudet, A Patel, C A Shaw, J R Lupski, A Gambin, S W Cheung, P Stankiewicz. 2013. "NAHR-Mediated Copy-Number Variants in a Clinical Population: Mechanistic Insights into Both Genomic Disorders and Mendelizing Traits." Genome Research 23 (9): 1395–1409.
IF=11.922; pkt MNiSW=50 (lista A)
-- pierwsze autorstwo współdzielone; **mój wkład: 40%**
Opis wkładu własnego: Moim głównym wkładem było przeprowadzenie analizy statystycznej, w szczególności analizy korelacji cech segmentalnych duplikacji (ang. Low Copy Repeats, LCR), oraz klastrów LCR z częstością występowania zmian liczby kopii (ang. Copy Number Variants, CNV) powstających *de novo*. Od strony informatycznej wyzwanie stanowiło zdefiniowanie i przygotowanie zbioru cech opisujących architekturę segmentalnych duplikacji

oraz dobór metod statystycznych do weryfikacji postawionej hipotezy. Po opracowaniu zbioru cech elementów LCR'ów oraz klastrów LCR, zaproponowałem schemat analizy statystycznej składający się z części eksploracyjnej (której celem było wstępne wyłonienie cech istotnie korelujących z częstością występowania *de novo* CNV) oraz części konfirmacyjnej, której istotą była budowa modelu najlepiej objaśniającego przyczyny różnic w częstościach powstawania zmian liczby kopii. Na potrzeby analizy eksploracyjnej wykorzystałem nieparametryczne testy porównujące cechy aktywnych (mediujących zdarzenia NAHR [ang. non-allelic homologous recombination]) i nieaktywnych regionów zawierających LCR'y. W fazie drugiej zbudowałem model regresji Poisson'a wykorzystujący wyłonione w fazie pierwszej cechy LCR'ów, bądź klastrów LCR'ów najlepiej objaśniający częstości zdarzeń NAHR w zależności od cech lokalnej architektury genomu. Na potrzeby pracy, przygotowałem również narzędzie do wizualizacji złożonej architektury genomowej w obrębie klastrów LCR. Program wykorzystuje algorytm ISCaas oraz pakiet miropeats. Dzięki integracji z przeglądarką genomową UCSC genome browser istnieje możliwość połączenia wygenerowanej przez miropeats ryciny z wizualizacją adnotacji genowych dostępnych w przeglądarce UCSC. Brałem udział w przygotowaniu manuskryptu oraz ostatecznej wersji artykułu.

[P2] P Dittwald*, **T Gambin***, C Gonzaga-Jauregui, C M B Carvalho, J R Lupski, P Stankiewicz, A Gambin. 2013. "Inverted Low-Copy Repeats and Genome Instability--a Genome-Wide Analysis." Human Mutation 34 (1): 210–20.
IF=4.601; pkt MNiSW=40 (lista A)
-- pierwsze autorstwo współdzielone; **mój wkład: 35%**
Opis wkładu własnego: Od strony informatycznej wyzwanie stanowiła integracja heterogenicznych danych o sekwencjach, adnotacjach genomowych oraz wynikach uzyskanych z sekwencjonowania i mikromacierzy. Na potrzeby pracy przeprowadziłem integrację danych dotyczących par odwróconych segmentalnych duplikacji z danymi o inwersjach, genach i ich znaczeniu klinicznym. Wspólnie z Piotrem Dittwaldem brałem udział w późniejszej analizie i interpretacji wyników, której wynikiem była identyfikacja fragmentów genomu oraz genów narażonych na nawracające inwersje. Brałem udział w przygotowaniu manuskryptu oraz ostatecznej wersji artykułu.

[P3] I M Campbell*, **T Gambin***, P Dittwald, C R Beck, A Shuvarikov, P Hixson, A Patel, A Gambin, C A Shaw, J A Rosenfeld, P Stankiewicz. 2014. "Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination." BMC Biology 12:74.
IF=6.779; pkt MNiSW=40 (lista A)
-- pierwsze autorstwo współdzielone; **mój wkład: 35%**
Opis wkładu własnego:
Zaproponowałem i zaimplementowałem rozwiązanie problemu rekonstrukcji kompletnych elementów typu HERV (ang. human endogenous retroviral elements), na podstawie dostępnych danych o fragmentach tych elementów. Opracowałem i zaimplementowałem efektywną procedurę wyznaczania uliniowienia par elementów typu HERV. Przyspieszenie procedury było możliwe dzięki zrównolegleniu obliczeń na wielu procesorach oraz zastosowaniu odpowiedniej filtracji wstępnej, która istotnie zredukowała liczbę potencjalnych par elementów typu HERV. Brałem udział w przygotowaniu manuskryptu oraz ostatecznej wersji artykułu. Konsultowałem budowę testu statystycznego wykorzystującego metodę Monte Carlo, służącego do sprawdzenia czy punty złamań rearanżacji są rozłożone w sposób niejednorodny w elementach typu HERV.

[P4] M Startek, P Szafranski, **T Gambin**, I M Campbell, P Hixson, C A Shaw, P Stankiewicz, and A Gambin. 2015. "Genome-Wide Analyses of LINE-LINE-Mediated Nonallelic Homologous Recombination." Nucleic Acids Research 43 (4): 2188–98.
IF=10.162; pkt MNiSW=40 (lista A)
--współautorstwo; **mój wkład: 15%**
Opis wkładu własnego: Przeprowadziłem analizę wzbogacenia okolic punktów złamań rearanżacji w motywy PRDM9. Konsultowałem konstrukcję modelu statystycznego opisującego wzbogacenie regionów narażonych na nawracające rearanżację w elementy typu LINE (ang. long interspersed nuclear element). Brałem udział w przygotowaniu manuskryptu oraz ostatecznej wersji artykułu.

[P5] T Gambin*, Z C Akdemir*, B Yuan, S Gu, T Chiang, C M B Carvalho, C Shaw, S Jhangiani, P M Boone, M K Eldomery, E Karaca, Y Bayram, A Stray-Pedersen, D Muzny, W-L Charng, V Bahrambeigi, J W Belmont, E Boerwinkle, A L Beaudet, R A Gibbs, J R Lupski. 2017. "Homozygous and Hemizygous CNV Detection from Exome Sequencing Data in a Mendelian Disease Cohort." Nucleic Acids Research 45 (4): 1633–48.
IF=10.162; pkt MNiSW=40 (lista A)
-- pierwsze autorstwo współdzielone; **mój wkład: 45%**
Opis wkładu własnego: Głównym wyzwaniem informatycznym było opracowanie narzędzia, pozwalającego na wykrywanie niewielkich zmian liczby kopii na podstawie danych z sekwencjonowania następnej generacji z większą czułością i precyzją niż istniejące metody. Zaprojektowałem i zaimplementowałem algorytm opisany w artykule. Zwiększenie czułości i precyzji w stosunku do innych algorytmów było możliwe dzięki zaproponowanym przeze mnie rozwiązaniom: (i) ograniczeniu zakresu poszukiwanych zmian do jednej klasy delecji (tzw. delecji homo-/hemi-zygotycznych), charakteryzujących się wyższym stosunkiem sygnału do szumu w stosunku do innych zmian liczby kopii; (ii) przeprowadzeniu jednoczesnej analizy głębokości pokrycia dla danych uzyskanych z sekwencjonowania dużej grupy pacjentów; (iii) wykorzystaniu dodatkowej informacji o wariantach pojedynczych nukleotydów otaczających delecję. Przeprowadziłem analizy porównawcze do innych narzędzi. Przygotowałem główną część manuskryptu dotyczącą wykorzystanych metod oraz uzyskanych wyników i ich interpretacji.

[P6] T Gambin, S N Jhangiani, J E Below, I M Campbell, W Wiszniewski, D M Muzny, J Staples, A C Morrison, M N Bainbridge, S Penney, A L McGuire, R A Gibbs, J R Lupski, E Boerwinkle. 2015. "Secondary Findings and Carrier Test Frequencies in a Large Multiethnic Sample." Genome Medicine 7 (1): 54.
IF=7.071; pkt MNiSW=40 (lista A)
--pierwsze autorstwo; **mój wkład: 80%**
Opis wkładu własnego: Główne wyzwanie informatyczne stanowiła jednoczesna analiza bardzo dużego zbioru danych z sekwencjonowania całoeksomowego (była to największa tego typu analiza przeprowadzona w owym czasie i obejmowała wyniki od ~11 tys. pacjentów). Moim istotnym wkładem było przeprowadzenie odpowiedniej kontroli jakości danych, zapewnienie spójności danych wejściowych oraz ujednolicenie sposobu adnotacji wariantów. Zaprojektowałem i zaimplementowałem system bazodanowy (zintegrowany z systemem **[T1]**) do przechowywania wariantów zoptymalizowany pod kątem zapytań SQL wykonywanych na potrzeby artykułu. Dzięki normalizacji danych system rozwiązał główny problem dotyczący spójności adnotacji pomiędzy próbkami. Opracowałem procedury importu danych z plików VCF (ang. Variant Call Format) oraz metody kontroli jakości, które umożliwiły wykrycie istotnych błędów w procedurach klinicznych oraz stanowiło motywację do pracy **[P7]**. Przygotowałem

główną część manuskryptu dotyczącą opisu opracowanych metod oraz uzyskanych wyników i ich interpretacji.

[P7] I M Campbell *, **T Gambin***, S N Jhangiani, M L Grove, N Veeraraghavan, D M Muzny, C A Shaw, R A Gibbs, E Boerwinkle, F Yu, J R Lupski. 2016. "Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses." Human Mutation 37 (3): 231–34.
IF=4.601; pkt MNiSW=40 (lista A)
--pierwsze autorstwo współdzielone; **mój wkład: 45%**
Opis wkładu własnego: Moim głównym wkładem było wykrycie skali problemu związanego z pomijaniem w analizach tzw. wariantów wielo-allelicznych. W pracy wykorzystałem dane zgromadzone w bazie danych przygotowanej przeze mnie na potrzeby artykułu **[P6]**. We współpracy z Ian'em Campbell'em przeprowadziłem analizę wpływu rozmiaru populacji na wzrost względnej liczby wariantów wielo-allelicznych. Przeprowadziłem analizę istniejących narzędzi i baz danych pod kątem wsparcia dla wariantów wielo-allelicznych. Brałem udział w przygotowaniu manuskryptu oraz ostatecznej wersji artykułu.

[P8] C Gonzaga-Jauregui, T Harel, **T Gambin**, M Kousi, L B Griffin, L Francescatto, B Ozes, E Karaca, S N Jhangiani, M N Bainbridge, K S Lawson, D Pehlivan, Y Okamoto, M Withers, P Mancias, Anne Slavotinek, Pamela J Reitnauer, Meryem T Goksungur, Michael Shy, Thomas O Crawford, M Koenig, J Willer, B N Flores, I Pediaditrakis, O Us, W Wiszniewski, Y Parman, A Antonellis, D M Muzny, N Katsanis, E Battaloglu, E Boerwinkle, R A Gibbs, J R Lupski. 2015. "Exome sequence analysis suggests that genetic burden contributes to phenotypic variability and complex neuropathy". Cell reports 12 (7): 1169-1183.
IF=8.282; pkt MNiSW=40 (lista A)
--współautorstwo (trzeci autor); **mój wkład: 15%**
Opis wkładu własnego: Mój wkład polegał na zaprojektowaniu, zaimplementowaniu i wykonaniu analiz statystycznych, które wykazały występowanie zjawiska zwiększonej liczby potencjalnie patogennych wariantów u grupy pacjentów chorych na złożone neuropatie, w stosunku do grupy kontrolnej. Dodatkowe wyzwanie informatyczne stanowiło zapewnienie spójności danych z różnych platform oraz kontrola jakości. Do rozwiązania tych problemów wykorzystałem bazę danych wariantów oraz procedury kontroli jakości opracowane w ramach pracy **[P6]**. Opracowałem część manuskryptu dotyczącą metod przygotowania danych oraz analiz statystycznych przeprowadzonych na potrzeby pracy. Brałem udział w przygotowaniu ostatecznej wersji artykułu.

[P9] M Wiewiórka, D Wysakowicz, M Okoniewski, **T Gambin**. 2017. "Benchmarking distributed data warehouse solutions for storing genomic variant information." Database: The Journal of Biological Databases and Curation, 2017:bax049.
IF=3.29; pkt MNiSW=40 (lista A)
-- ostatni autor (odpowiedzialny za sformułowanie problemu i nadzór nad projektem); **mój wkład: 30%**
Opis wkładu własnego: Moim wkładem było sformułowanie funkcjonalności hurtowni wariantów genetycznych. We współpracy z Markiem Wiewiórką oraz Dawidem Wysakowiczem opracowałem schemat hurtowni. Podczas implementacji wykorzystano mój pomysł na symulator danych testowych, skonstruowany w oparciu o rzeczywiste dane z bazy ExAC. Zaprojektowałem zestaw zapytań SQL, odpowiadających najbardziej popularnym przypadkom użycia. Brałem udział w testowaniu i interpretacji wyników wydajności poszczególnych konfiguracji rozproszonych silników i formatów danych oraz przygotowaniu manuskryptu i ostatecznej wersji artykułu.

[P10] A Hryhorzhevska, M Wiewiórka, M Okoniewski, **T Gambin**. 2017. "Scalable framework for the analysis of population structure using the next generation sequencing data." ISMIS 2017: Foundations of Intelligent Systems pp 471-480 (LNCS, volume 10352);
pkt MNiSW=15
-- ostatni autor (odpowiedzialny za sformułowanie problemu i nadzór nad projektem); **mój wkład: 30%**
Opis wkładu własnego: Moim wkładem było sformułowanie funkcjonalności skalowalnego narzędzia do obliczeń z zakresu genetyki populacyjnej umożliwiającego automatyczną kalibrację parametrów. Nadzorowałem implementację i ewaluację zaproponowanego rozwiązania. Uczestniczyłem w pisaniu manuskryptu oraz ostatecznej wersji artykułu.

[P11] U Lechowicz, **T Gambin**, A Pollak, A Podgorska, P Stawinski, A Franke, B-S Petersen, M Firczuk, M Oldak, H Skarżyński, R Płoski. 2017. "Iterative Sequencing and Variant Screening (ISVS) as a novel pathogenic mutations search strategy - application for *TMPRSS3* mutations screen". Scientific Reports 7:2543.
IF=4.259; pkt MNiSW=40 (lista A)
-- drugi autor; **mój wkład: 35%**
Opis wkładu własnego: Główne wyzwanie informatyczne stanowiła konstrukcja symulatora procedury Iterative Sequencing and Variant Screening (ISVS), służącej do identyfikacji wariantów patogennych oraz opracowanie miary oceny patogenności wariantów. Na potrzeby pracy zaprojektowałem oraz zaimplementowałem narzędzie do symulacji eksperymentów ISVS wraz z interfejsem użytkownika. Opracowałem również miarę oceniającą patogenność wariantów recesywnych, przy wykorzystaniu algorytmu klasyfikacji trenowanego na danych pochodzących z symulacji eksperymentu ISVS. W celu znalezienia najlepszego modelu klasyfikacji przeprowadziłem kalibrację i porównanie kilku algorytmów klasyfikacji pod nadzorem. Ponadto, wykonałem analizę stabilności wyników uzyskanych w eksperymencie ISVS dla różnych zestawów parametrów wejściowych. Moim wkładem było również opracowanie części manuskryptu dotyczącej konstrukcji symulatora ISVS oraz wykonanych za jego pomocą eksperymentów.

2.3 Wykaz opracowanych narzędzi informatycznych wchodzących w skład osiągnięcia oraz opis wkładu własnego

[T1] Autor: Tomasz Gambin
Data realizacji: 01.2013-30.12.2014
Tytuł: VariantAnalyzer - system bazodanowy do przechowywania i analizy danych z sekwencjonowania następnej generacji zaimplementowany i wdrożony przeze mnie na uczelni Baylor College of Medicine, na potrzeby projektu Centers for Mendelian Genomics. Kod źródłowy systemu jest dostępny pod adresem (https://github.com/BCM-Lupskilab/VariantAnalyzer). Część systemu umożliwiająca ko-segregację wariantów w rodzinach była prezentowana przeze mnie na konferencji American Sociaty of Human Genetics (ASHG) w roku 2013: (http://www.ashg.org/2013meeting/abstracts/fulltext/f130120890.htm).
Miejsce realizacji: Baylor College of Medicine, Houston, Texas, Stany Zjednoczone
Opis wkładu własnego: Mój wkład w zrealizowaniu tego osiągnięcia polegał na zaprojektowaniu i implementacji systemu.
Mój wkład 100%.

Wykorzystanie: System był wykorzystywany do analizy danych z NGS w pracach **[P5-P8]** oraz **[D1, D3-D5,D7-D14,D16-D21,D24-D27,D29-D32,D34-D37,D39-D43,D45-D47]** (wykaz dodatkowych publikacji, których jestem współautorem nie włączonych do osiągnięcia naukowego (D1-D52) znajduje się w ostatniej części autoreferatu).

[T2] Autor: Tomasz Gambin
Data realizacji: 05.2014-11.2016
Tytuł: HMZDelFinder - narzędzie do wykrywania homo i hemi-zygotycznych delecji dla danych z sekwencjonowania całoeksomowego (https://github.com/BCM-Lupskilab/HMZDelFinder) - opisane w pracy **[P5]**.
Miejsce realizacji: Baylor College of Medicine, Houston, Texas, Stany Zjednoczone oraz Instytut Informatyki, Politechnika Warszawska
Opis wkładu własnego: Mój wkład w zrealizowaniu tego osiągnięcia polegał na zaprojektowaniu i implementacji algorytmu.
Mój wkład: 100%
Wykorzystanie: System był wykorzystywany w pracy **[P5]** oraz w pracach **[D7,D12,D24]** .

[T3] Autorzy: Marek Wiewiórka, Dawid Wysakowicz, Michał Okoniewski, Tomasz Gambin
Data realizacji: 06.2016-05.2017
Tytuł: System do porównywania silników i formatów rozproszonych baz danych w kontekście przechowywania wariantów genetycznych (https://github.com/ZSI-Bio/variantsdwh) - opisany w pracy **[P9]**.
Miejsce realizacji: Instytut Informatyki, Politechnika Warszawska
Opis wkładu własnego: Mój wkład w zrealizowaniu tego osiągnięcia polegał na współudziale w projektowaniu schematu rozproszonej bazy danych. Byłem odpowiedzialny za projekt koncepcji symulatora danych, zdefiniowania przypadków użycia systemu oraz zapisania ich w postaci zapytań SQL. Zaimplementowałem część systemu wykorzystywaną do graficznej prezentacji wyników.
Mój wkład: 30%
Wykorzystanie: System był wykorzystywany w pracy **[P9]**.

[T4] **Autorzy:** Anastasiia Hryhorzhevska, Marek Wiewiórka, Michał Okoniewski, Tomasz Gambin
Data realizacji: 07.2016-01.2017
Tytuł: Popgen - system do kalibracji algorytmów grupowania i klasyfikacji danych genetycznych w celu identyfikacji odtworzenia struktury etnicznej badanej populacji (https://github.com/ZSI-Bio/popgen) - opisany w pracy **[P10]**.
Miejsce realizacji: Instytut Informatyki, Politechnika Warszawska
Opis wkładu własnego: Mój wkład w zrealizowaniu tego osiągnięcia polegał na zdefiniowaniu koncepcji systemu, współudziale w projektowaniu algorytmów oraz nadzorze nad ich implementacją.
Mój wkład: 30%
Wykorzystanie: System był wykorzystywany w pracy **[P10]**.

[T5] **Autor:** Tomasz Gambin
Data realizacji: 03.2015-05.2017

Tytuł: ISVS symulator - narzędzie do symulacji ekperymentów ISVS (http://zsibio.ii.pw.edu.pl/shiny/isvs/) - opisane w pracy **[P11]**.
Miejsce realizacji: Instytut Informatyki, Politechnika Warszawska
Opis wkładu własnego: Mój wkład w zrealizowaniu tego osiągnięcia polegał na zaprojektowaniu i implementacji algorytmu.
Mój wkład: 100%
Wykorzystanie: System był wykorzystywany w pracy **[P11]**.

2.4 Przyznane granty na prace badawcze, w ramach których powstały prace wchodzące w skład osiągnięcia naukowego

[G1] Centers for Mendelian Genomics, National Human Genome Research Institute grant 5U54HG006542. Rola w projekcie: wykonawca

[G2] Grant MNiSW Iuventus Plus pt. "Jednoczesna analiza wariantów pojedynczych nukleotydów oraz zmian strukturalnych uzyskanych z sekwencjonowania eksomowego lub celowanego". Rola w projekcie: kierownik

[G3] Grant NCN OPUS pt. "PerM-Cloud, Algorytmy i metody przetwarzania dużych zbiorów danych genomicznych w środowiskach chmur obliczeniowych na potrzeby personalizowanej medycyny". Kierownik grantu: Michał Okoniewski. Rola w projekcie: główny wykonawca

[G4] Grant dziekański pt "Przechowywanie i analiza danych genomowych z sekwencjonowania nowej generacji". Rola w projekcie: kierownik

3. Opis osiągnięcia naukowego

3.1 Wstęp

Dzięki odkryciu struktury DNA w latach 50 XX wieku, zsekwencjonowaniu pierwszego genomu ludzkiego w roku 2003 oraz rozwojowi technik sekwencjonowania następnej generacji (ang. Next Generation Sequencing, NGS), w ostatnim czasie nastąpił szybki rozwój w dziedzinach biologii molekularnej i genomiki. Ponadto, odkrycia nowych genów chorobowych, zrozumienie podstaw molekularnych wielu jednostek chorobowych, oraz możliwości wykorzystania nowych technologii sekwencjonowania w diagnostyce klinicznej doprowadziły do rewolucji we współczesnej medycynie, w której genomika zaczyna odgrywać coraz istotniejszą rolę.

Na genom człowieka składają się dwie kopie sekwencji, każda o długości ok. 3 miliardy nukleotydów, z których jedna jest odziedziczona od matki, a druga od ojca. Do celów analizy informatycznej sekwencja genomu reprezentowana jest jako ciąg symboli z alfabetu cztero literowego odpowiadającego zasadom wchodzącym w skład nukleotydów: adeninie (A), tyminie (T), cytozynie (C), guaninie (G). Fragmenty (pod-sekwencje) genomu zwane genami zawierają pełną informację o strukturze białek, które z kolei stanowią podstawowy budulec naszego organizmu oraz pełnią funkcje regulacyjne. Zmiany (warianty) powstałe w sekwencji DNA w wyniku mutacji genetycznych prowadzą zatem do zmian w budowie i funkcjonowaniu białek, co może skutkować rozwojem choroby. Badania genomu pacjenta, u którego podejrzewa się chorobę uwarunkowaną genetycznie, umożliwiają identyfikację wariantu, bądź grupy wariantów, które stanowią pierwotną przyczynę choroby. Postawienie diagnozy molekularnej coraz częściej otwiera możliwości zastosowania spersonalizowanych metod leczenia, a także daje pacjentowi

wiedzę na temat przewidywanych postępów choroby obserwowanych u innych osób z tą samą zmianą genetyczną. Do wykrywania wariantów obecnie stosuje się głównie technologie mikromacierzowe (pozwalające określić występowanie zmian w określonych pozycjach genomu) oraz techniki sekwencjonowania NGS (umożliwiające pełną analizę sekwencji genomu).

W moich badaniach zajmuję się tworzeniem narzędzi informatycznych na potrzeby genomiki, czyli nauki zajmującej się badaniem ogółu zjawisk zachodzących w genomach oraz zależności, jakie zachodzą pomiędzy nimi. Współcześnie, główny obszar zainteresowania w genomice stanowi informatyczna analiza danych pochodzących z wysokoprzepustowych metod badania genomu, w tym danych uzyskanych przy użyciu mikromacierzy oraz sekwenatorów następnej generacji. Przetwarzanie olbrzymich wolumenów wysoko-wymiarowych danych pochodzących z eksperymentów stanowi duże wyzwanie informatyczne. Z drugiej strony jednoczesna analiza genomów tysięcy pacjentów z bardzo wysoką rozdzielczością otwiera nowe możliwości poznawcze. Oprócz zwiększenia czułości i efektywności procedur diagnostycznych, pozwala m.in. na identyfikację nowych genów chorobowych, zrozumienie przyczyn powstawania mutacji oraz określenie zmienności genetycznej w różnych populacjach.

3.2 Specyfika danych genomowych

Złożona charakterystyka danych genomowych, w tym danych uzyskiwanych z wysokoprzepustowych metod genotypowania i sekwencjonowania, jest źródłem specyficznych wyzwań informatycznych, w szczególności związanych z:

- przetwarzaniem surowych danych uwzględniającym artefakty technologiczne wynikające z niedoskonałości metod eksperymentalnych;
- przetwarzaniem danych o znacznej liczbie wymiarów;
- jednoczesną analizą wielkich wolumenów danych.

3.2.1 Zaszumienie danych oraz występowanie systematycznych błędów

W analizie DNA człowieka większość danych genomowych analizuje się w kontekście tzw. genomu referencyjnego, czyli publicznie dostępnej sekwencji DNA o długości ok. 3 miliardów nukleotydów. Podstawowym celem eksperymentów biologicznych jest wykrywanie zmian (wariantów) występujących w genomie osoby badanej w stosunku do sekwencji referencyjnej. Mogą być to zarówno zmiany pojedynczych nukleotydów jak i większe rearanżacje, w tym delecje (ubytki materiału genetycznego), duplikacje (naddatki materiału genetycznego), inwersje (odwrócenie fragmentów genomu), translokacje (przeniesienie materiału genetycznego z jednej części genomu do innej). O ile zjawiska genomowe mają charakter dyskretny (np. występowanie lub brak wariantu), to surowe dane eksperymentalne, które wykorzystuje się do wykrywania tych zmian, pochodzą z pomiarów zjawisk fizycznych mających charakter ciągły (np. poziomy świecenia barwników fluorescencyjnych). Brak możliwości bezpośredniego odczytu sekwencji genomu oraz niedoskonałości technologiczne istniejących rozwiązań wprowadzają potencjalne źródło szumu. Inne błędy w danych wynikają między innymi z faktu, że materiał genetyczny, wykorzystywany w eksperymentach, pozyskiwany jest jednocześnie z wielu komórek (z których każda może zawierać trochę inną sekwencję genomową) i poddawany jest złożonej obróbce chemicznej. Znaczna część artefaktów technologicznych ma charakter błędów systematycznych (tj. powtarzających się w kolejnych eksperymentach). Przy tworzeniu narzędzi informatycznych na potrzeby analiz genomowych, należy zwrócić szczególną uwagę na modelowanie i uwzględnienie źródeł błędów systematycznych w celu minimalizacji ich wpływu na wyniki.

3.2.2 Bardzo duża liczba wymiarów

Dane uzyskane w technologii hybrydyzacji porównawczej do mikromacierzy (ang. array-based Comparative Genomic Hybridization, aCGH) bądź sekwencjonowania NGS cechuje bardzo wysoka liczba wymiarów, która wynika z jednoczesnej analizy sekwencji DNA w różnych obszarach genomu. Przykładowo, surowe dane z aCGH zawierają informacje na temat względnej liczby kopii (pomiędzy genomem osoby badanej a genomem referencyjnym) dla kilkuset tysięcy fragmentów DNA. Z kolei, surowe dane z sekwencjonowania pojedynczego pacjenta zawierają informacje na temat setek milionów, a nawet setek miliardów odczytów sekwencyjnych odpowiadających fragmentom DNA pochodzącym z różnych części genomu. W wyniku analizy odczytów sekwencyjnych otrzymujemy listę zawierającą do kilku milionów wariantów dla każdego pacjenta. Liczba wymiarów (różnych wariantów genetycznych) obserwowanych podczas jednoczesnej analizy dużych populacji sięga dziesiątek, a nawet setek milionów.

3.2.3 Jednoczesna analiza wielkich wolumenów danych

Wysoka wymiarowość danych bezpośrednio przekłada się na ich duży rozmiar, w szczególności w odniesieniu do jednoczesnej analizy danych pochodzących od dziesiątek lub setek tysięcy osób badanych. W związku z tym praktycznie na każdym etapie przetwarzania danych genomowych pojawiają się problemy wynikające z dużej złożoności obliczeniowej i pamięciowej algorytmów i struktur danych. Z drugiej strony, jednoczesna analiza dużych zbiorów danych: (i) umożliwia usunięcie systematycznych błędów (patrz punkt 3.2.1); (ii) ułatwia proces kontroli jakości i pozwala na szybką identyfikację elementów odstających; (iii) usprawnia proces identyfikacji rzadkich zmian genomowych, które stanowią główną przyczynę chorób uwarunkowanych genetycznie.

3.3 Wkład własny w rozwój rozwiązań informatycznych na potrzeby przetwarzania danych genomowych

Głównym celem moich prac badawczych był rozwój nowych i adaptacja istniejących narzędzi informatycznych na potrzeby analizy danych genomowych, które następnie znajdowały zastosowanie w rozwiązywaniu rzeczywistych problemów biologicznych i medycznych.

Rozwiązywanie problemów biomedycznych z wykorzystaniem danych genomowych jest procesem wieloetapowym. Rozpoczyna go analiza zagadnienia biomedycznego, w której kluczowym elementem jest przyswojenie przez badacza powiązanej z danym problemem wiedzy genetycznej, medycznej i biochemicznej. Następnym krokiem jest abstrakcja problemu prowadząca do zdefiniowania niezbędnych elementów (danych, wiedzy, etapów przetwarzania), które będą wykorzystane w rozwiązaniu informatycznym. Na dalszym etapie, poszukuje się gotowych narzędzi informatycznych lub bioinformatycznych, które mogą zostać wykorzystane w kolejnych krokach przetwarzania. Zadanie to wymaga dokładnego zbadania i porównania możliwości poszczególnych narzędzi, a także ewentualnej adaptacji, jeżeli rozwiązanie nie oferuje w pełni wymaganej funkcjonalności. Alternatywnie, mamy możliwość opracowania i wprowadzenia własnej metody. W takim wypadku przygotowane przez nas narzędzie musi zostać poddane ewaluacji i porównaniu z konkurencyjnymi rozwiązaniami, jeżeli takie istnieją.

Ze względu na różnorodność zagadnień przetwarzania danych jakie pojawiają się podczas rozwiązywania problemów biomedycznych, w moich badaniach wykorzystywałem różne klasy rozwiązań informatycznych (podsumowane w tabeli 3.3.1). Bardziej dokładny opis podejmowanych przeze mnie problemów biomedycznych oraz sposobów ich rozwiązania przy

wykorzystaniu opracowanych przeze mnie narzędzi znajduje się w sekcji 3.4.

Tabela 3.3.1: Zagadnienia z obszaru informatyki rozwijane na potrzeby realizacji prac wchodzących w skład osiągnięcia naukowego.

Zagadnienia z obszaru informatyki	**Publikacje i narzędzia wchodzące w skład osiągnięcia naukowego**	
	Analiza, zastosowanie i adaptacja istniejących narzędzi	**Opracowanie własnych rozwiązań**
Optymalizacja złożoności i przyspieszenie obliczeń na danych genomowych	P1-P8, P11, T1,T2, T5	
Metody przetwarzania rozproszonego	P9, P10, T3, T4	P10, T4
Metody statystyczne	P1-P8, T2	P3, P4, P6, P8
Metody przetwarzania sygnałów	P5, T2	P5, T2
Metody uczenia maszynowego	P1, P10, P11, T3, T4	
Metody wizualizacji danych	P1-P5, P7, T2	P5, T2

3.3.1 Optymalizacja złożoności i przyspieszenie obliczeń na danych genomowych

Genom referencyjny stanowi punkt odniesienia dla większości danych genomowych. Ponieważ materiał genetyczny człowieka składa się z osobnych fragmentów DNA zwanych chromosomami, każdej lokalizacji w genomie możemy przyporządkować odpowiednią nazwę chromosomu (chr1, chr2, … , chr22, chrX, chrY) oraz pozycję na tym chromosomie. Dzięki procedurze mapowania (określania położenia na genomie referencyjnym) możliwa jest integracja i wspólna analiza danych pochodzących z wielu eksperymentów dotyczących tych samych, bądź różnych rodzajów zjawisk genomowych. Wydajne przetwarzanie tego typu danych wymaga wykorzystania algorytmów i struktur danych umożliwiających szybkie operacje na przedziałach genomowych, w szczególności przecięcia, wyszukiwania części wspólnych/rozłącznych, dużych zbiorów punktów bądź przedziałów.

W opracowywanych przeze mnie rozwiązaniach szczególny nacisk był położony na aspekt optymalizacji złożoności obliczeniowej tworzonych algorytmów. Redukcja czasu wykonania analizy wielokrotnie pozwoliła na weryfikację większej liczby hipotez oraz dokładniejszą kalibrację parametrów wejściowych, w porównaniu z metodami alternatywnymi. W kontekście przetwarzania danych genomowych wydajna implementacja narzędzi, wykonujących liczne operacje na dużych zbiorach przedziałów, była możliwa dzięki wykorzystaniu struktur typu NClist (ang. Nested Containment List) zaimplementowanych w postaci pakietu R o nazwie "IRanges" oraz nakładki dla danych genomowych, t.j. pakietu "GenomicRanges". Złożoność obliczeniowa zapytań (np. przecięcie zbiorów przedziałów) wykorzystujących tę strukturę wynosi O(n + logN), gdzie N to liczba przedziałów, natomiast n to

liczba zwróconych wyników. Tego typu podejście jest nawet kilkusetkrotnie szybsze od rozwiązań indeksowania stosowanych w bazach takich jak mySQL (B-drzewo) czy Postgres (R-drzewo) (Alekseyenko and Lee 2007; Lawrence et al. 2013) i było przeze mnie szeroko wykorzystywane w pracach **[P1-P8]**. Na potrzeby kolejnych publikacji poszerzałem zakres funkcjonalności biblioteki GenomicRanges, dodając m.in.: (i) metodę wyliczającą liczbę przecięć z przedziałami jednej z listy pierwszej dla każdego przedziałów z drugiej listy; (ii) metodę adnotacji obszarów genomowych przy wykorzystaniu listy genów.

Nieodłącznym elementem tworzenia oprogramowania na potrzeby przetwarzania danych genomowych jest ciągła analiza możliwości potencjalnego zrównoleglenia obliczeń. W zależności od rodzaju wykonywanej w danym kroku analizy, zrównoleglenia można dokonać na różnym poziomie granularności, m.in: na poziomie próbek, chromosomów, pojedynczych elementów genomowych. Zrównoleglenie okazało się szczególnie istotne w przypadku uliniawiania (algorytmem Smith'a-Waterman'a) dużej liczby par elementów typu HERV (ang. human endogenous retroviral elements), które przeprowadziłem na potrzeby realizacji pracy **[P3]**. Podobnie, zrównoleglenie poszczególnych kroków przetwarzania w zaimplementowanym przeze mnie algorytmie HMZDelFinder **[T2]** pozwoliło znacząco zredukować całkowity czas wykonania obliczeń. W pracy **[P11]**, zrównoleglenie, opracowanego przeze mnie, algorytmu symulacji umożliwiło bardziej precyzyjną ewaluację stabilności zaproponowanego rozwiązania dzięki zwiększeniu liczby iteracji.

Aby umożliwić modularyzację procesu przetwarzania danych z sekwencjonowania, w środowisku bioinformatycznym, został wypracowany standardowy zestaw formatów plików [FASTQ (Cock et al. 2010), Binary Alignment Map (BAM) (H. Li et al. 2009), Variant Call Format (VCF) (Danecek et al. 2011)], które są wykorzystywane do przechowywania danych wytworzonych w kolejnych etapach analizy. Według tego schematu, ostateczna lista wariantów znalezionych u osoby badanej przechowywana jest w plikach VCF, zawierających informację o lokalizacji każdego wariantu oraz jego jakości pozwalającej oszacować prawdopodobieństwo, czy wykryta zmiana jest prawdziwa. Dodatkowo, aby usprawnić proces priorytetyzacji i interpretacji znalezionych wariantów, w plikach VCF umieszcza się tzw. adnotacje wariantów, czyli zestaw informacji, pochodzących z zewnętrznych źródeł danych, szczegółowo opisujących możliwe konsekwencje wystąpienia danego wariantu. O ile przetwarzanie plików VCF, dla pojedynczych, bądź niewielkich grup pacjentów, nie stanowi problemu, to jednoczesna analiza tysięcy przypadków staje się bardzo niewydajna. Utrudniony i czasochłonny jest również proces readnotacji wariantów, dokonywanej w celu uaktualnienia informacji pochodzących z zewnętrznych źródeł danych. W celu usprawnienia analiz wielopróbkowych, na potrzeby moich badań **[P6, P7]**, opracowałem i zaimplementowałem schemat relacyjnej bazy danych wykorzystywanej do składowania i efektywnej eksploracji danych o wariantach genetycznych. Istotną różnicą w stosunku do standardowego podejścia (stosowanego w plikach VCF), było odseparowanie informacji o wystąpieniu wariantów u poszczególnych pacjentów od meta danych (adnotacji wariantów). Zastosowanie znormalizowanej struktury tabel, umożliwiło utworzenie efektywnych metod aktualizacji adnotacji wariantów.

3.3.2 Metody przetwarzania rozproszonego

Głównym problemem, tradycyjnych, zcentralizowanych modeli przetwarzania jest brak skalowalności. Oznacza to, że algorytmy działają i wykonują operacje w zadowalającym czasie dopóki rozmiar danych nie przekracza pewnego poziomu. Szybki przyrost danych genomowych, coraz częściej prowadzi do sytuacji, w której jednoczesna analiza całości zbioru danych przestaje być możliwa.

Problem braku skalowalności w tradycyjnym modelu przetwarzania, dotyczy zarówno algorytmów służących do analizy jak i narzędzi przeznaczonych do składowania danych.

Zaobserwowane ograniczenia w opracowanym przeze mnie rozwiązaniu do przechowywania wariantów genetycznych, opartym na modelu relacyjnym, skłoniły mnie do badań nad możliwościami wykorzystania metod przetwarzania rozproszonego oraz narzędzi z obszaru Big Data w genomice. Przeprowadzone przeze mnie, we współpracy z zespołem, systematyczne porównanie wydajności różnych konfiguracji rozproszonych formatów danych (ORC, Parquet, Kudu) oraz silników zapytań (Spark SQL, Hive, Presto, Impala) pozwoliło potwierdzić istotną przewagę tego typu rozwiązań w stosunku do tradycyjnych, relacyjnych baz danych **[P9, T3]**. Ponadto, wyniki analizy wskazały konkretne konfiguracje narzędzi rozproszonych pozwalające uzyskać najmniejsze czasy odpowiedzi dla różnych rodzajów zapytań (odpowiadającym różnym obszarom zastosowań analiz genomowych).

Narzędzia wykorzystujące rozproszone silniki obliczeń (takie jak Apache Spark, Flink) znajdują większe zastosowanie w przetwarzaniu danych z NGS. Najbardziej popularne potoki przetwarzania (ang. pipelines) do wykrywania wariantów (Van der Auwera et al. 2013), są obecnie reimplementowane przy użyciu narzędzi z ekosystemu Hadoop. W pracy **[P10]** przedstawiony został, opracowany przeze mnie we współpracy z zespołem, nowy, skalowalny potok przetwarzania **[T4]** dedykowany na potrzeby analiz z zakresu genetyki populacyjnej, wykorzystujący rozproszone wersje algorytmów zaimplementowanych w Apache Spark, służących m.in. do redukcji wymiarów, klasyfikacji, grupowania. Porównanie z konkurencyjnymi metodami potwierdziło zdecydowanie wyższą wydajność naszego rozwiązania.

3.3.3 Metody statystyczne

Typowym podejściem do weryfikacji hipotez biomedycznych jest przeprowadzenie odpowiedniego testu statystycznego. W implementacjach opracowywanych przeze mnie narzędzi do analizy danych genomowych, szczególnie często wykorzystywałem klasę testów i statystyk nieparametrycznych (m.in. testy Kolmogorov–Smirnov, Mann-Whitney-Wilcoxon, Fisher-exact, korelacje rangowe Spearmana) **[P1-P8, P11]**. Podejście nieparametryczne pozwala uniknąć dodatkowych założeń dotyczących rozkładów, które w przypadku danych genomowych są często nieznane, bądź trudne do ustalenia, ze względu na dużą liczbę czynników wpływających na charakterystykę danych (zobacz również sekcję 3.2.1). Wyjątkiem były sytuacje, w których rodzina rozkładów wynikała bezpośrednio z typu danych. Przykładowo, w pracy **[P1]** wykorzystałem regresję Poisson'a do modelowania rozkładu częstości wystąpień zdarzeń genomowych.

Poza zastosowaniem gotowych testów i miar statystycznych, przy weryfikacji części hipotez badawczych wystąpiła konieczność opracowania własnych statystyk, i miar oceny. W szczególności, na potrzeby pracy **[P6]**, zaimplementowałem eksperyment randomizacyjny, pozwalający na oszacowanie ryzyka wystąpienia choroby recesywnej u dziecka, którego rodzice pochodzą z określonej populacji. W pracy **[P8]** zaprojektowałem i zaimplementowałem test permutacyjny wykorzystujący statystyki nieparametryczne z testu Mann-Whitney-Wilcoxon'a, który pozwolił wykazać występowanie dodatkowych mutacji (ang. mutational load) w grupie pacjentów w stosunku do grupy kontrolnej. Zaproponowana przez zespół, którego byłem członkiem, statystyka w pracy **[P4]**, potwierdziła że pomiędzy elementami typu LINE (ang. long interspersed nuclear element), częściej dochodzi do rearanżacji genomowych niż w innych, losowych miejscach genomu. W pracy **[P3]**, opracowany został nowy test, wykorzystujący metodę Monte Carlo, który pozwolił wykazać nielosowe grupowanie punktów złamań rearanżacji położonych w elementach typu HERV.

Ważnym aspektem analiz statystycznych w genomice, jest odpowiednie przygotowanie danych przed wykonaniem testu. W moich pracach, szczególną uwagę poświęciłem problemowi przygotowania danych na potrzeby tzw. testów asocjacyjnych. Służą one do wykrywania

nowych zależności genotypowo-fenotypowych, w szczególności do identyfikacji nowych genów chorobowych. Tego typu metody były początkowo rozwijane na potrzeby analiz GWAS (ang. Genome-Wide Association Studies) wykonywanych dla wariantów częstych, uzyskanych z mikromacierzy (Rentería, Cortes, and Medland 2013). Rozwój technologii NGS pozwolił na uwzględnienie w analizach asocjacyjnych, także wariantów rzadko występujących w populacji, które odgrywają znacznie ważniejszą rolę w patogenezie chorób uwarunkowanych genetycznie, niż warianty częste. Na potrzeby analizy rzadkich wariantów zaproponowano w literaturze nowe rodzaje testów asocjacyjnych (tzw. testy agregacyjne, m.in. testy typu Burden (Morgenthaler and Thilly 2007), Sequence Kernel Association Test (SKAT) (Wu et al. 2011)), które pozwalają na jednoczesne porównywanie częstości całych grup wariantów położonych w zadanych obszarach genomu (np. genach, grupach genów, itd.). W większości metod wykorzystywane są modele regresji pozwalające uwzględnić różnego rodzaju czynniki potencjalnie zakłócające wyniki (ang. confounding factors). Czynnikiem, istotnie wpływającym na genotyp jest pochodzenie etniczne osób badanych. Opracowane przez zespół, którego byłem członkiem, narzędzie **[P10, T4]**, pozwala na szybkie i dokładne określenie pochodzenia etnicznego, umożliwiając tym samym usprawnienie procesu przygotowania danych na potrzeby testów asocjacyjnych. Podobnie, zaproponowane w pracy **[P9]** skalowalne rozwiązania służące do przechowywania danych o wariantach, pozwalają na efektywną eksplorację danych w kontekście badań kliniczno-kontrolnych (ang. case-control) oraz wydajną implementację fragmentów testów asocjacyjnych bezpośrednio w języku SQL.

3.3.4 Metody przetwarzania sygnałów

Do wykrywania zmian strukturalnych na podstawie danych z mikromacierzy lub NGS wykorzystuje się techniki z dziedziny przetwarzania sygnałów, m.in. takie jak normalizacja, segmentacja, progowanie. Dla danych mikromacierzowych, do celów segmentacji sygnału log2ratio najpowszechniej wykorzystywana jest metoda CBS (ang. Circular Binary Segmentation) (Olshen et al. 2004). W pracy **[P5]** zaproponowałem sposób wykorzystania metody CBS do analizy danych o częstości B-allelicznej (ang. B-allele frequency, BAF) uzyskanych z danych NGS, w celu wykrycia regionów genomowych wykazujących braki heterozygotyczności (ang. Absence of Heterozygosity, AOH).

Powyższy algorytm został przeze mnie wykorzystany w konstrukcji nowej metody wykrywania homo- i hemi-zygotyczych delecji (HMZDelFinder) **[T2, P5]** na podstawie danych z NGS. W kontekście konkurencyjnych metod (Fromer et al. 2012; Krumm et al. 2012; Jiang et al. 2015; Packer et al. 2016; Guo et al. 2014), innowacyjny charakter tego narzędzia, polega na: (a) nowym podejściu do wykrywania elementów odstających w danych o głębokości pokrycia, które umożliwiło osiągnąć wysoką czułość rozwiązania (b) jednoczesnym wykorzystaniu dwóch ortogonalnych źródeł danych (o głębokości pokrycia oraz częstości B-allelicznej), co pozwoliło znacząco zwiększyć precyzję. Przeprowadzone przeze mnie testy na publicznie dostępnych oraz wewnętrznych zbiorach danych potwierdziły przewagę algorytmu HMZDelFinder w stosunku do innych rozwiązań, w szczególności w kontekście identyfikacji trudno wykrywalnych, krótkich delecji.

3.3.5 Metody uczenia maszynowego

Metody klasyfikacji (zarówno bez nadzoru jak i pod nadzorem) znajdują zastosowanie w wielu obszarach analiz genomowych. W moich pracach, dotyczących analizy danych z NGS **[P6, P8, P9, P11]** korzystałem z opisanych w literaturze rozwiązań opartych na metodach klasyfikacji, służących m.in. do: (i) określania jakości wariantów (Van der Auwera et al. 2013); (ii) przewidywania patogenności wariantów (X. Liu et al. 2016); (iii) przewidywania nowych genów chorobowych (X. Liu et al. 2016; Lek et al. 2016).

Poza wykorzystaniem wyników gotowych narzędzi, w moich badaniach stosowałem metody klasyfikacji i grupowania do rozwiązania nowych problemów biomedycznych. W pracy **[P1]**, we współpracy z Piotrem Dittwaldem, użyliśmy metodę wykorzystująca hierarchiczny algorytm grupowania, w celu identyfikacji klastrów segmentalnych duplikacji (powtarzających się długich, fragmentów DNA o wysokim stopniu podobieństwa) w genomie referencyjnym, co pozwoliło zdefiniować obszary narażone na nawracające rearanżacje. Wyzwanie informatyczne stanowiła kalibracja algorytmu (m.in. dobór miary odległości pomiędzy klastrami, wybór liczby klastrów), której celem była identyfikacja parametrów metody zapewniających najlepsze dopasowanie do danych o opisanych w literaturze klastrach segmentalnych duplikacji. Innym przykładem wykorzystania technik uczenia maszynowego w analizie danych genomowych jest praca **[P10]**, w której, we współpracy zespołem, dokonałem porównania kilku skalowalnych (zaimplementowanych w Apache Spark) metod grupowania [hierarchiczne, k-średnich, EM (ang. Expectation-Maximization)] oraz wybranych metod klasyfikacji [maszyny wektorów nośnych (ang. Support Vector Machine, SVM), drzew decyzyjnych, lasów losowych] pod kątem ich skuteczności w przewidywaniu pochodzenia etnicznego. Dodatkowym wyzwaniem było opracowanie procedury kalibracji parametrów przetwarzania wstępnego. W pracy **[P11]** opisałem wybór i kalibrację modelu klasyfikacji służącego do dyskryminacji patogennych i niepatogennych wariantów recesywnych z wykorzystaniem danych pochodzących z eksperymentu ISVS (ang. Iterative Sequencing and Variant Screening).

3.3.6 Metody wizualizacji danych

W moich pracach, metody wizualizacji stosowałem na potrzeby prezentacji wyników analiz. W pracach **[P1-P4, P7]**, wykorzystałem zmodyfikowane przeze mnie elementy biblioteki quantsmooth (Eilers and de Menezes 2004) z repozytorium Bioconductor do prezentacji wyników w formie tzw. ideogramów reprezentujących strukturę cytogenetyczną chromosomów. Modyfikacje polegały m.in. na dodaniu możliwości prezentacji danych dla wersji genomu hg19, oraz funkcji umożliwiającej wizualizację dowolnych adnotacji genomowych. W pracy **[P2]** wystąpiła konieczność wizualizacji większej liczby różnych adnotacji genomowych na jednej rycinie. Do tego celu został wykorzystany pakiet CIRCOS (Krzywinski et al. 2009). W pracy **[P1]** wyzwanie stanowiła wizualizacja złożonej architektury miejsc narażonych na nawracające rearanżację. Na potrzeby prezentacji wyników wykorzystałem algorytm ISCaas oraz pakiet miropeats (Parsons 1995), który zintegrowałem z wizualizacją wybranych adnotacji generowanych za pomocą przeglądarki genomowej "UCSC genome browser" (Karolchik, Hinrichs, and James Kent 2007).

Własne metody wizualizacji danych zaproponowałem w pracy **[P5]**. Opracowane funkcje służą do generacji wykresów obrazujących delecje oraz regiony AOH wykryte za pomocą narzędzia HMZDelFinder. Wykresy wykrytych delecji, prezentują szczegółową informację na temat zmian w głębokości pokrycia w próbce z delecją na tle danych o głębokości pokrycia w pozostałych próbkach. Wizualna inspekcja wykresów umożliwia identyfikację znalezisk fałszywie pozytywnych, które nie zostały odfiltrowane w ramach automatycznej procedury kontroli jakości.

3.4 Zastosowania w biologii i medycynie

3.4.1 Analiza danych o zmianie liczby kopii DNA z mikromacierzy CGH w celu identyfikacji niestabilnych regionów genomu ludzkiego

Technologia aCGH umożliwia detekcję szczególnego rodzaju wariantów genetycznych, czyli niezrównoważonych zmian liczby kopii DNA (CNV) z wysoką rozdzielczością. Zmiany liczby

kopii to przedziały w sekwencji genomowej, które różnią się liczbą kopii między genomami dwóch osób i powstają w wyniku delecji bądź duplikacji części genomu. Eksperyment aCGH polega na hybrydyzacji pofragmentowanych i oznaczonych barwnikami fluorescencyjnymi DNA genomu pacjenta oraz genomu osoby zdrowej, które następnie są hybrydyzowane do sond oligonukleotydowych umieszczonych na mikromacierzy. Analiza intensywności poziomów świecenia pozwala określić względną liczbę kopii fragmentów DNA u pacjenta w stosunku do osoby zdrowej.

Moje badania prowadzone podczas studiów doktoranckich doprowadziły do opracowaniem nowych algorytmów służących do tworzenia własnych projektów mikromacierzy. We współpracy z Instytutem Matki i Dziecka (IMID) w Warszawie (http://naukawpolsce.pap.pl/aktualnosci/news,398593,polska-macierz-wykrywa-wady-genomu.html) oraz uczelnią Baylor College of Medicine (BCM) w Houston w Stanach Zjednoczonych przygotowałem pierwszą na świecie diagnostyczna mikromacierz CGH do wykrywania CNV z eksonową rozdzielczością. Wstępne wyniki uzyskane przy wykorzystaniu tej mikromacierzy w laboratoriach Baylor Genetics (BG) zostały opublikowane w pracy (Boone et al. 2010). Projekty mikromacierzy opracowanych przeze mnie we współpracy z BCM weszły również do oferty firmy Agilent (https://www.genomeweb.com/arrays/agilent-makes-baylor-designed-cgh-arrays-available-constitutional-and-cancer-res). **Podsumowując, moja praca doktorska była skoncentrowana na rozwiązaniu problemów związanych z projektowaniem mikromacierzy oraz przetwarzaniem surowych danych aCGH, w celu wykrywania CNV. Prace, które wchodzą w skład przedstawionego tutaj osiągnięcia naukowego, nie podejmują bezpośrednio tej tematyki, ale stanowią jej logiczną kontynuację.**

Po obronie doktoratu zająłem się tworzeniem narzędzi informatycznych na potrzeby analizy miejsc w genomie ludzkim szczególnie narażonych na nawracające rearanżacje. Jednym z głównych mechanizmów powstawania nawracających rearanżacji jest proces tzw. nie-allelicznej rekombinacji homologicznej (ang. non-allelic homologous recombination, NAHR). W wyniku NAHR, do rearanżacji genomowych dochodzi częściej w charakterystycznych miejscach w genomie, otoczonych przez sekwencje DNA, które wykazują w stosunku do siebie znaczące podobieństwo (homologie). W genomie ludzkim wyróżniamy różne klasy powtarzających się sekwencji, m.in.: segmentalne duplikacje, czyli długie proste lub odwrócone zduplikowane fragmenty DNA wykazujący wysoki stopień podobieństwa (>90%), oraz różnego rodzaju elementy transpozonowe jak elementy LINE czy HERV.

Segmentalne duplikacje (ang. Segmental Duplications) określane również jako powtórzenia o niskiej liczbie kopii (ang. Low-Copy Repeats, LCRs) obejmują od 4-5% ludzkiego genomu (Bailey et al. 2001). We wcześniejszych pracach pokazano, że w wyniku NAHR w obrębie segmentalnych duplikacji mogą powstawać punkty złamań nawracających duplikacji i delecji (Lupski 1998) (Stankiewicz and Lupski 2002). Do zmian liczby kopii DNA pomiędzy prostymi (skierowanymi w tym samym kierunku) segmentalnymi duplikacjami (ang. Directly oriented paralogous LCRs, DP-LCRs) dochodzi dwa razy częściej niż w innych miejscach w genomie (J. Li et al. 2012). Pojawiła się również publikacja sugerująca, że pewne cechy segmentalnych duplikacji (np. długość sekwencji homologicznej) mogą mieć wpływ na częstość występowania nawracających rearanżacji pomiędzy daną parą sekwencji zduplikowanych (P. Liu et al. 2012).

W pracy **[P1]**, we współpracy z Piotrem Dittwaldem, przeprowadziliśmy całogenomową analizę informatyczną architektury segmentalnych duplikacji. W celu identyfikacji regionów narażonych na nawracające rearanżacje w pierwszym kroku dokonaliśmy agregacji znanych sekwencji LCR w większe klastry wykorzystując algorytm grupowania hierarchicznego. Parametry algorytmu grupowania, zostały dobrane w oparciu o analizę wybranych znanych regionów mikroduplikacyjnych i mikrodelecyjnych. W ten sposób wyznaczyliśmy 105

regionów ograniczonych przez pary klastrów DP-LCR, z czego ponad połowę stanowiły regiony uprzednio powiązane z syndromomi mikrodelecyjnymi i mikroduplikacyjnymi. Następnie, wykorzystując dane z mikromacierzy CGH, zebrane w laboratoriach diagnostycznych BCM od ponad 25 tys. pacjentów zidentyfikowaliśmy nawracające zmiany liczby kopii zlokalizowane regionach ograniczonych przez klastry DP-LCR. W przypadku 190 CNV zostało potwierdzone, że nie występowały one u rodziców pacjenta, czyli powstały *de novo*, najpewniej w wyniku procesu NAHR. Informację o różnicach w częstościach występowania zmian *de novo* w różnych regionach wykorzystałem do identyfikacji cech par klastrów DP-LCR takich jak długość LCR, odległość pomiędzy klastrami, stopień podobieństwa pomiędzy parami klastrów [ang. fraction matching], częstość występowania nukleotydów GC (ang. GC content) oraz częstości występowania wybranych motywów w największym stopniu sprzyjających powstawaniu rearanżacji. W pierwszej części analizy wykorzystałem statystyczne testy nieparametryczne w celu porównania każdej cechy pomiędzy grupą aktywnych i nieaktywnych regionów flankowanych przez DP-LCR. Następnie, przeprowadziłem analizę korelacji pomiędzy cechami klastrów DP-LCR a częstościami występowania *de novo* CNV w regionach flankowanych przez te klastry. Do analizy wykorzystałem nieparametryczne testy korelacji rangowej Spearman'a (w ramach analizy eksploracyjnej) oraz model regresji Poisson'a (analiza konfirmacyjna). W wyniku przeprowadzonej analizy udało się określić nieopisane wcześniej w literaturze, nowe cechy elementów DP-LCR znacznie zwiększające ryzyko wystąpienia zmiany chromosomalnej. Podsumowując, rezultaty uzyskane w pracy stanowią istotny wkład do zrozumienia natury NAHR. Co więcej mogą zostać wykorzystane przy budowie narzędzi służących do predykcji lokalizacji w genomie, które są szczególnie narażone na nawracające rearanżacje.

W podobny sposób jak zgodnie skierowane segmentalne duplikacje (DP-LCRs) pośredniczą w powstawaniu delecji i duplikacji, tak odwrotnie skierowane segmentalne duplikacje (ang. Inverse Paralogous LCRs, IP-LCRs) mogą przyczyniać się do powstawania inwersji. Ponieważ wykrywanie inwersji jest zadaniem dużo bardziej skomplikowanym niż detekcja delecji i duplikacji, wiedza o skali tego zjawiska jest wciąż niepełna. W pracy **[P2]**, we współpracy z zespołem, przedstawiłem wyniki analizy informatycznej, której zadaniem była identyfikacja w genomie ludzkim miejsc flankowanych przez IP-LCRs, czyli regionów potencjalnie narażonych na występowanie nawracających inwersji. Przeprowadzona analiza wykazała, że takie regiony obejmują aż 12% genomu człowieka i prawie 1000 genów może ulegać uszkodzeniom w wyniku nawracających inwersji.

Wypracowane przeze mnie metody analizy statystycznej oraz narzędzia informatyczne służące do badania regionów flankowanych przez DP-LCRs oraz IP-LCRs przygotowane w pracach **[P1]** i **[P2]** zostały wykorzystane w pracy **[P3]**, w której wykazaliśmy, że oprócz segmentalnych duplikacji, również elementy typu HERV mogą pośredniczyć w procesie NAHR. Na potrzeby analizy, opracowałem algorytm rekonstrukcji kompletnych elementów typu HERV oraz metodę identyfikacji par elementów typu HERV, które potencjalnie mogą pośredniczyć w powstawaniu rearanżacji. Następnie, wspólnie z Ian'em Campbell'em, przeszukaliśmy diagnostyczną bazę danych BCM w celu identyfikacji CNV, których końce były flankowane przez wybrane uprzednio pary elementów HERV. Dla części z CNV udało się określić dokładne punkty złamań rearanżacji i zlokalizować ich położenie w elementach HERV przy wykorzystaniu niezależnej metody biologicznej (sekwencjonowania metodą Sangera). Dla par elementów typu HERV, które obejmowały więcej niż jedną rearanżację zaobserwowaliśmy, że punkty złamań różnych rearanżacji rozmieszczone są nierównomiernie w obszarze danego elementu HERV tworząc wyraźne skupiska. Aby wykazać nierównomierność opracowaliśmy test statystyczny wykorzystujący metodę Monte Carlo, który potwierdził lokalne koncentracje punktów złamań. Dodatkowa analiza wykazała, że w okolicach zgrupowanych punktów złamań częściej występują charakterystyczne krótkie sekwencje (motywy) związane, w obrębie których dochodzi do rekombinacji (ang. recombination hotspot motif).

Moje kolejne badania nad niestabilnością genomu ludzkiego **[P4]**, dotyczyły analizy roli elementów typu LINE w procesie NAHR. Po dokonaniu identyfikacji par LINE mogących pośredniczyć w NAHR, przeszukano bazę kilkuset tysięcy CNV zgromadzonych w bazie BCM dzięki czemu znaleziono ponad 500 przypadków, w których CNV mogły być potencjalnie mediowane przez elementy typu LINE. Aby lepiej zrozumieć jakie cechy elementów repetytywnych sprzyjają zwiększonej niestabilności genomu w danym regionie, Michał Startek, we współpracy z zespołem, którego byłem członkiem, opracował statystykę, pozwalającą na oszacowanie wzbogacenia punktów złamań obserwowanych CNV w elementy typu LINE. Mój udział w powstaniu nowej miary oceny polegał na identyfikacji problemu związanego z koniecznością uwzględnienia artefaktów wynikających z konstrukcji mikromacierzy, takich jak brak sond w obrębie nie-unikatowych fragmentów genomu, takich jak elementy typu LINE. Wartości statystyki zostały wyznaczone osobno dla par LINE o różnej długości oraz różnym poziomie podobieństwa sekwencji, co pozwoliło stwierdzić, że to właśnie poziom podobieństwa najmocniej koreluje z częstością występowania zmian liczby kopii.

3.4.2 Analiza danych z sekwencjonowania następnej generacji

W przeciwieństwie do mikromacierzy, które umożliwiają detekcję zmian w wybranych regionach genomu, sekwencjonowanie następnej generacji (NGS) umożliwia odczyt pełnej sekwencji genomowej, a tym samym wykrywanie wariantów genetycznych zarówno na poziomie pojedynczych nukleotydów (ang. Single Nucleotide Variants, SNV) jak i większych zmian strukturalnych (w tym CNV). Sekwencjonowanie dokonuje się na tzw. płytkach reakcyjnych (ang. flow cells) na których hybrydyzowane są miliony, krótkich fragmentów DNA, pochodzące z genomu osoby badanej. W procesie sekwencjonowania przez syntezę (ang. sequencing by synthesis) sekwencja każdego z fragmentów zostaje odczytana i zapisana w celu dalszego przetwarzania. Surowe wyniki można dalej wykorzystać do asemblacji *de novo*, czyli składania nowych genomów lub do wykrywania wariantów genetycznych w przypadku gdy genom referencyjny gatunku jest znany (jak w przypadku genomu człowieka). W celu obniżenia kosztów, dość często zamiast sekwencjonowania pełnego genomu stosuje się tzw. sekwencjonowanie celowane, które obejmuje wszystkie (sekwencjonowane eksomowe) bądź wybraną grupę sekwencji kodujących białka (genów). W celu identyfikacji wariantów patogennych surowe dane analizuje się w tzw. potoku przetwarzania (ang. pipeline), na który składają się moduły odpowiadające za: (i) proces mapowania i uliniowienia odczytów do genomu referencyjnego; (ii) oznaczanie zduplikowanych odczytów oraz proces rekalibracji zmapowanych sekwencji; (iii) wykrywanie wariantów oraz ocenę ich jakości; (iv) adnotacje wariantów. Standardowe potoki przetwarzania są tworzone przede wszystkim do wykrywania i analizy zmian typu SNV, natomiast do identyfikacji zmian strukturalnych stosuje się osobne narzędzia informatyczne.

Po obronie doktoratu, odbyłem dwuletni staż (lata 2013-2014) w laboratorium prof. Jima Lupskiego na uczelni Baylor College of Medicine w Houston w Stanach Zjednoczonych, gdzie brałem udział w projekcie badawczym "Centers for Mendelian Genomics (CMG)" ufundowanym przez Amerykański Instytut Zdrowia (National Health Institute, NIH) **[G1]**. Projekt ten stanowił kontynuację słynnego projektu Human Genome Project, który zakończył się w roku 2003 odczytaniem pierwszej sekwencji ludzkiego genomu. Celem projektu CMG było rozpoznanie przyczyn molekularnych leżących u podłoża chorób o dziedziczeniu Mendlowskim, dla których do tej pory nie zostały zidentyfikowane geny chorobowe, t.j. geny w których mutacje prowadzą do choroby. Moja rola w projekcie CMG obejmowała nadzór nad potokiem przetwarzania danych z sekwencjonowania oraz kontrolą jakości uzyskanych wyników. Ponadto byłem odpowiedzialny za rozwój środowiska bazodanowego oraz graficznych interfejsów służących do przechowywania wyników, wstępnego przetwarzania i zarządzania danymi oraz

odpowiadałem za koordynację prac analityków/biologów/lekarzy zajmujących się interpretacją danych. Zajmowałem się również opracowywaniem nowych narzędzi informatycznych służących m. in. do identyfikacji strukturalnych, w tym regionów wykazujących brak heterozygotyczności (AOH) oraz zmian liczby kopii (CNV) na podstawie danych z sekwencjonowania całoeksomowego. W ramach badań przeprowadzałem też analizy, których celem była identyfikacja nowych genów chorobowych, w których wykorzystywałem zarówno istniejące jak i tworzyłem własne narzędzia informatyczne służące do ko-segregacji wariantów w rodzinach, czy wykonywania testów asocjacyjnych.

3.4.2.1 System do przechowywania i analizy danych z sekwencjonowania NGS

W okresie trwania pierwszej fazy projektu CMG (lata 2012-2015) w centrum sekwencjonowania Human Genome Sequencing Center (HGSC) należącym do BCM wygenerowano sekwencję eksomowe (ang. Whole Exome Sequencing, WES) dla ponad 5,000 pacjentów. Analiza tak wielkiego zbioru danych stanowiła duże wyzwanie informatyczne. Na potrzeby projektu opracowałem system bazodanowy (VariantAnalyzer), który został przeze mnie wdrożony i jest nadal wykorzystywany podczas analizy danych z projektu CMG **[T1]**. System wykonałem w oparciu o wzorzec projektowy Model View Controller, przy użyciu frameworku web2py zaimplementowanego w języku Python. Dzięki integracji z systemem klasy Laboratory Information Management System (LIMS) wykorzystywanym w HGSC do śledzenia wyników wstępnej analizy danych z sekwencjonowania, VariantAnalyzer zapewnia automatyczną synchronizację i aktualizację danych o sekwencjach eksomowych. Po każdorazowej synchronizacji system automatycznie aktualizuje lokalną bazę częstości wariantów. System umożliwia jednoczesny dostęp dla wielu użytkowników, z których każdy ma dostęp ograniczony do próbek wchodzących w skład podprojektów, których analizą się zajmuje. Oprócz dostępu do zaadnotowanej listy wariantów, użytkownik w ramach systemu VA może przeglądać zmiany strukturalne (AOH, CNV) uzyskane na podstawie analizy danych z WES. Inne funkcje systemu umożliwiają wyszukiwanie wszystkich wystąpień wariantów w lokalnej bazie dla zadanej listy genów.

3.4.2.2 Detekcja AOH oraz CNV z WES

Narzędzia do wykrywania CNV z WES opierają się na założeniu, że liczba kopii DNA dla zadanego fragmentu genomu jest skorelowana dodatnio z głębokością pokrycia w tym regionie. Oznacza to, że głębokość pokrycia w przypadku delecji będzie mniejsza, a w przypadku duplikacji większa w stosunku do regionów o niezmienionej liczbie kopii. Ze względu na znaczące wahania pokrycia, wynikające m.in. z artefaktów technologicznych związanych z procesem wzbogacenia, przed właściwą detekcją zmian liczby kopii dokonuje się wstępnego przetworzenia danych w celu zwiększenia stosunku sygnału do szumu. Do tego celu wykorzystuje się uśrednioną głębokość pokrycia dla każdego z regionów należących do wzbogacenia (tj. dla kolejnych eksonów), która jest następnie poddawana normalizacji względem głębokości pokrycia uzyskanej w innych próbkach. Jednoczesna wielo-próbkowa analiza pozwala na zniwelowanie systematycznych fluktuacji sygnału oraz identyfikację potencjalnych zmian liczby kopii, czyli regionów w danej próbce, dla których głębokość pokrycia istotnie różni się od głębokości w pozostałych próbkach. Mimo wykorzystania różnorodnych technik modelowania statystycznego, w tym ukrytych modeli Markowa, analizy składowych głównych (Fromer and Purcell 2014; Krumm et al. 2012), których celem jest redukcja szumu , możliwości detekcji zmian obejmujących mniej niż trzy eksony pozostają nadal bardzo ograniczone (de Ligt et al. 2013).

W pracy **[P5]** przedstawiłem opracowany przeze mnie algorytm HMZDelFinder **[T2]**, umożliwiający wykrywanie homo- i hemi-zygotycznych delecji pojedynczych eksonów. Możliwość wykrycia tego typu zmian ze zwiększoną rozdzielczością w stosunku do

tradycyjnych metod wykrywania CNV, ma szczególne znaczenie w przypadku analizy pacjentów, u których podejrzewana jest choroba recesywna, bądź choroba sprzężona z chromosomem X. W przeciwieństwie do innych metod detekcji CNV, HMZDelFinder korzysta z informacji o absolutnej wartości pokrycia (zamiast wartości znormalizowanej) dzięki czemu jest w stanie zidentyfikować regiony w badanej próbce, w których brakuje pokrycia bądź jest ono stosunkowo niskie. Z drugiej strony, informacja o głębokości pokrycia w innych próbkach pozwala na odfiltrowanie regionów nie informatywnych, dla których niskie pokrycie obserwowane jest w istotnym fragmencie populacji próbek kontrolnych. Innym ważnym, zaproponowanym przeze mnie elementem algorytmu, który pozwolił kilkukrotnie zwiększyć jego precyzję, jest weryfikacja czy potencjalna delecja homo- lub hemi-zygotyczna znajduje się obszarze wykazującym braki heterozygotyczności. Na potrzeby detekcji regionów AOH, opracowałem algorytm wykorzystujący informację o częstości B-allelicznej (BAF) którą uzyskałem na podstawie analizy plików VCF, jako stosunek liczby odczytów odpowiadających allelowi alternatywnemu do całkowitej głębokości pokrycia dla kolejnych wariantów. Następnie odpowiednio przetransformowany wektor BAF jest poddawany procedurze segmentacji wykorzystującej algorytm CBS (Olshen et al. 2004) w celu identyfikacji spójnych fragmentów genomu o podobnej gęstości występowania wariantów heterozygotycznych, w tym regionów nie zawierających wariantów heterozygotycznych, czyli AOH.

Podsumowując, dzięki połączeniu dwóch źródeł danych (o głębokości pokrycia oraz częstości B-allelicznej wariantów) udało mi się opracować algorytm zapewniający znacznie wyższą czułość i precyzję wykrywania delecji homo- i hemi-zygotycznych w porównaniu z innymi narzędziami, na co dowodem są wyniki eksperymentów przeprowadzone na publicznie dostępnym zbiorze danych 1000 Genomes (Aebersold and Malmstroem 2013) oraz na zbiorze danych wygenerowanym w ramach projektu CMG.

3.4.2.3 Jednoczesna analiza dużych zbiorów WES

Współpraca z HGSC (w szczególności z prof. Eric'iem Boerwinkle) w ramach grantu CMG zaowocowała pomysłem na projekt, którego celem była analiza dużego zbioru WES (> 11,000 sekwencji) pod kątem występowania w próbkach tzw. znalezisk przypadkowych (ang. incidental findings) oraz nosicielstwa mutacji w znanych genach odpowiedzialnych za choroby recesywne (ang. recessive carriers). Znaleziska przypadkowe, określane również mianem znalezisk wtórnych (ang. secondary findings) to warianty, mające istotne znaczenie kliniczne, nie powiązane z głównym rozpoznaniem klinicznym, które stanowiło przyczynę wskazania do badania genetycznego. Amerykańskie towarzystwo ACMG (ang. American College of Medical Genetics and Genomics) w roku 2015 opublikowało listę 56 genów w których znalezione wszystkie patogenne mutacje powinny być raportowane niezależnie od podstawowego wskazania na badanie (Smith et al. 2015). Podobnie jak znaleziska przypadkowe, nosicielstwo wariantów patogennych w genach odpowiedzialnych za choroby recesywne może stanowić klinicznie istotną informację o pacjencie.

Na tle poprzednich prac o podobnej tematyce, publikacja **[P6]** wyróżnia się wykorzystaniem dużo większego i bardziej zróżnicowanego (jeżeli chodzi o strukturę etniczną) zbioru danych sekwencji całoeksomowych. Od strony informatycznej wyzwanie stanowiło odpowiednie przygotowanie danych, w tym połączenie danych z dwóch projektów (CMG oraz ARIC - Atherosclerosis Risks in Communities) i uwspólnienie procedur kontroli jakości, a także stworzenie narzędzi pozwalających na efektywną eksplorację połączonych danych. Tradycyjny sposób przechowywania listy wariantów w plikach VCF nie spełniał oczekiwań jeżeli chodzi o wydajność i swobodę przetwarzania w przypadku tak dużego wolumenu danych. W związku z tym, na potrzeby projektu zaprojektowałem oraz zaimplementowałem dedykowaną relacyjną bazę danych, która umożliwia sprawne zarządzanie oraz przeszukiwanie zgromadzonych wariantów i zapewniła znaczną poprawę wydajności przetwarzania w stosunku do analizy

plików VCF. Opracowana baza danych pozwoliła mi udzielić rzetelnych odpowiedzi na postawione w pracy **[P6]** pytania dotyczące m.in. częstości występowania znanych lub przewidywanych wariantów patogennych oraz ich rozkładu w różnych grupach etnicznych.

Doświadczenia zebrane podczas pracy nad publikacją **[P6]**, opracowana baza danych i stworzone narzędzia informatyczne pozwoliły mi na identyfikację istotnego problemu związanego z występowaniem pozycji, dla których w populacji występują więcej niż dwa różne allele (ang. multi-allelic sites). Ze względu na brak przystosowania narzędzi informatycznych do problemu wielo-alleliczności wariantów oraz trudności w ich przetwarzaniu, znaczna część publicznie dostępnych zbiorów danych niepoprawnie reprezentowała zmiany znajdujące się na tych pozycjach. W pracy **[P7]**, we współpracy z Ian'em Campbell'em dokonałem szczegółowej analizy problemu, pokazując konsekwencje, jakie wiążą się z nieprawidłowym traktowaniem pozycji wielo-allelicznych. Zaprezentowałem również dane wskazujące na szybki wzrost wielkości problemu wraz z przyrostem liczby próbek. Ostatecznie dokonałem analizy najpopularniejszych narzędzi informatycznych w kontekście ich wsparcia dla wariantów wielo-allelicznych.

Wspomniana powyżej baza danych wariantów została również wykorzystana w pracy **[P8]**, w której celem było określenie wpływu dodatkowych mutacji (poza główną mutacją wskazaną w rozpoznaniu jako przyczyna choroby) występujących u pacjentów ze złożonymi neuropatiami. Na potrzeby porównania liczby niesynonimicznych mutacji u pacjentów względem próbek kontrolnych, opracowałem i zaimplementowałem permutacyjny test statystyczny wykorzystujący wyniki nieparametrycznych testów Mann-Whitney-Wilcoxon'a. Wyniki analizy przeprowadzonej dla dwóch niezależnych populacji pacjentów ze złożonymi neuropatiami potwierdziły występowanie zwiększonego tzw. ładunku mutacji (ang. mutational load), czyli dodatkowych mutacji poza mutacją patogenną występujących w genach odpowiedzialnych za neuropatie.

3.4.2.4 Skalowalne metody analizy danych z NGS

Po powrocie ze stażu podoktorskiego w roku 2015 wznowiłem pracę w Instytucie Informatyki na Politechnice Warszawskiej. Wspólnie z dr. hab Michałem Okoniewskim oraz mgr. inż. Markiem Wiewiórką utworzyliśmy nieformalną grupę badawczą ZSI-Bio w Zakładzie Systemów Informatycznych Instytutu Informatyki, której głównym celem jest rozwój skalowalnych narzędzi i rozwiązań chmurowych na potrzeby genetyki i genomiki. Obrany kierunek badań został wybrany nieprzypadkowo i wynika z moich wcześniejszych doświadczeń i problemów jakie pojawiały się podczas realizacji grantu CMG. Dotyczyły one m.in. problemów związanych z niską wydajnością przetwarzania surowych danych, a także składowania i współdzielenia danych pomiędzy różnymi zespołami. Implementacja algorytmów analizy danych genomowych z wykorzystaniem rozproszonych silników obliczeniowych oraz rozproszonych baz danych, pozwala rozwiązać znaczną część z tych problemów. Dalszy rozwój zespołu, m.in. o nowych dyplomantów i doktorantów był możliwy dzięki finansowaniu jakie uzyskaliśmy w ramach grantów **[G2]**, **[G3]** oraz **[G4]**.

Centralnym problemem we współczesnej genomice, jest kwestia przechowywania danych w sposób, który zapewnia możliwości ich efektywnej eksploracji. O ile opisane powyżej, opracowane przeze mnie rozwiązania wykorzystujące relacyjne bazy danych sprawdziły się w projekcie CMG, to dla większych wolumenów (rzędu dziesiątek/setek tysięcy sekwencji całoeksomowych bądź całogenomowch) ich wydajność jest niewystarczająca. Stąd powstał pomysł na opracowanie hurtowni danych wariantów genetycznych w pełni zintegrowanej ze środowiskiem do obliczeń rozproszonych.

Praca **[P9]** opisuje, opracowany, przeze mnie i zespół, system **[T3]** służący do porównywania wydajności różnych konfiguracji rozproszonych silników (Spark SQL, Hive, Presto, Impala) i formatów bazodanowych (ORC, Parquet, Kudu). Na potrzeby systemu

zaproponowałem schemat prototypowej bazy wariantów genetycznych w oparciu o strukturę gwiazdy często wykorzystywaną przy konstrukcji hurtowni danych. Centralna tabela (tabela faktów) zawierała informacje o genotypach (wystąpieniach wariantów u poszczególnych pacjentów) natomiast pozostałe tabele (tabele wymiarów) zawierały informację na temat adnotacji wariantów, lokalizacji genów/eksonów, a także informację dotyczące osób badanych (grupa etniczna, fenotyp). Dodatkowo, we współpracy z Dawidem Wysakowiczem, opracowałem symulator do generacji syntetycznych danych testowych. Dzięki wykorzystaniu rzeczywistych informacji o częstości wariantów w różnych grupach etnicznych z bazy ExAC (Lek et al. 2016) oraz informacji o adnotacjach z bazy dbNSFP (X. Liu et al. 2016) charakterystyka wygenerowanych wariantów jest zbliżona do rzeczywistej. Symulator został użyty do generacji listy wariantów odpowiadającej danym z 50,000 sekwencji całoeksomowych. W ramach testów opracowałem zestaw kilkunastu zapytań SQL, które odpowiadają najczęstszym przypadkom wykorzystania bazy wariantów do celów klinicznych (analiza pojedynczych pacjentów) jak i badawczych (złożone zapytania analityczne, np. pozwalające na przeprowadzenie testów asocjacyjnych). Na podstawie wyników przeprowadzonych eksperymentów, wspólnie z zespołem, przedstawiliśmy rekomendacje na temat użyteczności konkretnych konfiguracji silników bazodanowych i formatów danych do budowy produkcyjnej wersji hurtowni wariantów genetycznych. Należy podkreślić, że opracowany system **[T3]** jest łatwo rozszerzalny, a proces testowania został w pełni zautomatyzowany, dlatego będzie mógł być w przyszłości wykorzystywany do testowania i porównywania wydajności nowych rozwiązań bazodanowych.

Wielo-próbkowe dane z sekwencjonowania następnej generacji, dzięki wykorzystaniu dodatkowej informacji genetycznej, którą niosą ze sobą rzadkie warianty, umożliwiają analizę struktury etnicznej badanej populacji z rozdzielczością znacznie większą niż pozwalały na to dane uzyskane z technik mikromacierzowych. Analiza struktury etnicznej na danych z NGS w dużych badaniach populacyjnych stanowi ważny element kontroli jakości pozwalający wychwycić niespójności w metadanych oraz odstające obserwacje, mogące sugerować błędy sekwencjonowania. Ponadto, informacja o strukturze etnicznej badanej populacji jest często wykorzystywana przy budowaniu modeli regresji na potrzeby testów asocjacyjnych dla rzadkich wariantów (Santorico and Hendricks 2016). Niestety, tradycyjne metody przetwarzania nie radzą sobie z analizą znacznej liczby wysoko-wymiarowych danych z NGS. W literaturze brakuje również systematycznego porównania czułości różnych algorytmów klasyfikacji i grupowania oraz metod pozwalających na automatyczną kalibrację kolejnych kroków wstępnego przetwarzania, t.j. selekcji wariantów, selekcji atrybutów po redukcji cech wykonanej przy użyciu algorytmów PCA (ang. Principal Component Analysis) lub MDS (ang. Multidimensional Scaling).

W pracy **[P10]**, we współpracy z Anastasiią Hryhorzhevską, Markiem Wiewiórką oraz Michałem Okoniewskim, opracowałem dwa systemy przeznaczone do analiz populacyjnych **[T4]**. Pierwszy z nich, zaimplementowany w języku R wykorzystuje elementu pakietu SNPRelate (Zheng et al. 2012) i pozwala zautomatyzować proces optymalizacji parametrów oraz wyboru algorytmów uczenia maszynowego zapewniających największą jakość grupowania i klasyfikacji. System został przetestowany na rzeczywistych, publicznie dostępnych danych z projektu 1000 Genomes. Drugie narzędzie zostało zaimplementowane w rozproszonym środowisku Apache Spark co pozwoliło wielokrotnie zwiększyć wydajność przetwarzania i umożliwiło przeprowadzenie jednoczesnej analizy kilkudziesięciu tysięcy próbek w akceptowalnym wymiarze czasowym. W stosunku do konkurencyjnych implementacji rozproszonych (O'Brien et al. 2015), zaproponowane przez nasz zespół rozwiązanie osiągnęło wyraźnie lepsze wyniki jakości grupowania.

3.4.2.5 Nowa metoda wykrywania patogennych wariantów w genach powiązanych z chorobami dziedziczonymi w modelu autosomalnie recesywnym

Praca **[P11]** przedstawia nową metodę nazwaną ISVS służącą do wykrywania wariantów patogennych w genach odpowiedzialnych za choroby dziedziczone w modelu autosomalnie recesywnym. Wykorzystanie tej metody dla dużej grupy pacjentów, u których podejrzewa się chorobę wywołaną mutacją bi-alleliczną w genie powiązanym z chorobą recesywną, pozwala znacząco obniżyć koszty diagnostyczne w porównaniu z analizą obejmującą sekwencjonowanie całej populacji. Aby zainicjalizować procedurę ISVS potrzebna jest identyfikacja znanych mutacji patogennych u jednego z pacjentów. Następnie dla wszystkich znalezionych potencjalnie patogennych mutacji wykonuje się badania przesiewowe przy wykorzystaniu PCR (ang. Polymerase Chain Reaction) u wszystkich pozostałych pacjentów. Dla osób u których w wyniku PCR znaleziono pojedynczą mutację przeprowadza się następnie pełne sekwencjonowanie danego genu w celu identyfikacji drugiej zmiany. W ten sposób wykrywane są nowe potencjalnie patogenne mutacje, które następnie poddaje się analizie PCR w całej populacji. Proces powtarzany jest dopóki w procesie sekwencjonowania odnajdywane są nowe potencjalnie patogenne zmiany.

Aby zweryfikować skuteczność działania metody oraz jej czułość na parametry wejściowe, opracowałem symulator ISVS wraz z interfejsem graficznym, który zaimplementowałem w języku R przy wykorzystaniu frameworku shiny **[T5]**. Dodatkowo, opracowałem algorytm pozwalający na dyskryminację wariantów patogennych od wariantów łagodnych na podstawie informacji o częstości współwystępowania danego wariantu z innym wariantem potencjalnie patogennym. W celu wyłonienia najlepszego klasyfikatora dokonałem kalibracji i porównania kilku modeli klasyfikacyjnych (wykorzystujących lasy losowe, regresję logistyczna, maszynę wektorów wspierających, drzewa decyzyjne). Podsumowując, opracowany przeze mnie system pozwala ocenić użyteczność metody ISVS dla konkretnego zastosowania na podstawie zdefiniowanych przez użytkownika założeń (m.in. dotyczących skumulowanej częstości wariantów patogennych w interesującym użytkownika genie), oraz umożliwia ocenę patogenności wykrytych wariantów.

3.5 Podsumowanie

Moje badania po doktoracie koncentrowały się na rozwoju metod i narzędzi informatycznych, niezbędnych w analizie danych genomowych. Mój wkład polegał na opracowaniu: (i) narzędzi służących do wydajnego przetwarzania, analizy statystycznej i wizualizacji danych genomowych, które umożliwiły poszerzenie wiedzy na temat przyczyn niestabilności genomu i powstawania zmian typu CNV wynikających z lokalnej architektury genomu; (ii) metod pozwalających na bardziej skuteczne, w stosunku do istniejących rozwiązań, wykrywanie zmian strukturalnych (CNV, AOH) przy wykorzystaniu danych z NGS; (iii) modeli relacyjnych systemów bazodanowych oraz skalowalnych narzędzi do składowania i eksploracji danych z NGS pozwalających na jednoczesną analizę danych z dużych projektów sekwencjonowania genomów.

Przy projektowaniu i implementacji narzędzi bioinformatycznych, korzystałem z umiejętności i doświadczeń jakie nabyłem podczas studiów magisterskich i doktoranckich w Instytucie Informatyki, PW. W szczególności, szeroko wykorzystywałem wiedzę z zakresu projektowania relacyjnych i rozproszonych baz danych, konstrukcji wielowarstwowych systemów informatycznych oraz metod eksploracji danych, wnioskowania statystycznego i sztucznej inteligencji.

Opracowane przeze mnie rozwiązania informatyczne zostały z sukcesem wykorzystane w analizach dziesiątek tysięcy wyników uzyskanych z mikromacierzy CGH oraz

sekwencjonowania następnej generacji przyczyniając się do zwiększenia skuteczności metod diagnostycznych oraz odkryć nowych genów chorobowych. Rezultaty tych badań zostały opublikowane w kilkudziesięciu renomowanych czasopismach (w tym 3 publikacje w czasopiśmie *Cell* **[D5,D41,D46]** oraz jedna w *Nature Genetics* **[D29]**). Artykuły te nie wchodzą w skład prezentowanego osiągnięcia naukowego, ale są z nim ściśle powiązane. Przykładowo, opracowany przeze mnie system VariantAnalyzer **[T1]** oraz powiązana z nim baza danych wariantów, a także narzędzia do analizy AOH, analizy kosegregacji wariantów, wykonywania testów asocjacyjnych były wykorzystywane w pracach **[D1, D3-D5,D7-D14,D16-D21,D24-D27,D29-D32,D34-D37,D39-D43,D45-D47]**. Analiza zbioru danych CMG wykonana za pomocą algorytmu HMZDelFinder **[T2]** umożliwiła identyfikację patogennych CNV opisanych m.in. w pracach **[D7,D12,D24]**.

Ponadto, doświadczenia zebrane podczas stażu podoktorskiego w BCM, w tym wiedza o budowie potoków przetwarzania oraz opracowane narzędzia informatyczne wykorzystuje obecnie w analizie danych polskich pacjentów, sekwencjonowanych w Zakładzie Genetyki Medycznej Instytutu Matki i Dziecka (ZGM IMID). Ze względu, na szybki rozwój technologii sekwencjonowania, malejące koszty i związany z tym szybki przyrost danych, na znaczeniu coraz bardziej zyskują skalowalne rozwiązania wykorzystujące rozproszone silniki obliczeniowe oraz rozproszone bazy danych. Dlatego, jest wysoce prawdopodobne, że narzędzia **[T3,T4]**, a także nowe rozwiązania do analizy CNV i SNV implementowane w ramach grantu **[G2,G3]**, pozwolą na dalszą poprawę skuteczności diagnostycznej metod wykorzystujących NGS oraz umożliwią nowe odkrycia w zakresie patogenezy chorób uwarunkowanych genetycznie.

4 Pozostałe osiągnięcia naukowo - badawcze

4.1 Wskaźniki bibliometryczne

Liczba publikacji w czasopismach z listy A (JCR)	72 (w tym 61 po doktoracie)		
Liczba publikacji w czasopismach z listy B	3 (w tym 1 po doktoracie)		
Artykuły z innych czasopism	3		
Rozdziały z książek	1 (w tym 1 po doktoracie)		
Materiały konferencyjne (WoS)	3 (w tym 2 po doktoracie)		
Łączna liczba punktów ministerialnych	2817 *		
Sumaryczny Impact Factor	482 *		
Wskaźnik	**Google Scholar**	**Scopus (z wyłączeniem autocytowań)**	**Web of Science (z wyłączeniem autocytowań)**
Liczba cytowań	1804	1190	1100

Indeks Hirsha	24	21	20

* - obliczone na podstawie danych zgromadzonych w repozytorium Instytutu Informatyki (http://www.ii.pw.edu.pl/ii_pol/Instytut-Informatyki/Organizacja-Instytutu/ZSI/Pracownicy)

4.2 Nagrody i wyróżnienia

09.2016: Laureat XI konkursu na stypendia naukowe Ministerstwa Nauki i Szkolnictwa Wyższego dla wybitnych młodych naukowców.

09.2016: Indywidualna nagroda Rektora Politechniki Warszawskiej: nagroda pierwszego stopnia za dorobek naukowy w latach 2014-2015.

07.2014: ASHG/Charles J. Epstein Trainee Award for Excellence in Human Genetics Research – półfinalista.

09.2013: Indywidualna nagroda Rektora Politechniki Warszawskiej: nagroda pierwszego stopnia za dorobek naukowy w latach 2011-2012.

03.2011: Stypendium Fundacji Nauki Polskiej START (edycja 2011).

04.2010: Stypendium Fundacji Nauki Polskiej START (edycja 2010).

12.2010: Stypendium otrzymane w ramach konkursu dla doktorantów organizowanego przez Centrum Studiów Zaawansowanych PW (CAS/16/POKL), współfinansowane ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

04.2009: Mazowieckie Stypendium Doktoranckie "Mazovia", współfinansowane ze środków Unii Europejskiej w ramach poddziałania 8.2.2 PO KL pn. „Rozwój nauki – rozwojem regionu – stypendia i wsparcie dla mazowieckich doktorantów".

2008 – 2012: Stypendia za wyniki na studiach doktoranckich.

10.2009: Nagroda za najlepszy plakat na konferencji "International Multiconference on ComputerScience and Information Technology" w Mrągowie.

06.2007: Obrona pracy magisterskiej z wyróżnieniem.

4.3 Spis publikacji uzyskanych po doktoracie nie wchodzących w skład osiągnięcia naukowego

[D1] Gambin T, Yuan B, Bi W, Liu P, Rosenfeld JA, et al. Identification of novel candidate disease genes from *de novo* exonic copy number variants. Genome Med. 2017 Sep 21;9(1):83. IF=7.071; pkt MNiSW=40

[D2] Zhang J, Gambin T, Yuan B, Szafranski P, Rosenfeld JA, et al. Haploinsufficiency of the E3 ubiquitin-protein ligase gene TRIP12 causes intellectual disability with or without autism

spectrum disorders, speech delay, and dysmorphic features. Hum Genet. 2017 Apr;136(4):377-386. IF=4.637 ; pkt MNiSW=35

[D3] Bekheirnia MR, Bekheirnia N, Bainbridge MN, Gu S, Coban Akdemir ZH, et al. Whole-exome sequencing in the molecular diagnosis of individuals with congenital anomalies of the kidney and urinary tract and identification of a new causative gene. Genet Med. 2017 Apr;19(4):412-420. IF=7.329 ; pkt MNiSW=40

[D4] Eldomery MK, Coban-Akdemir Z, Harel T, Rosenfeld JA, Gambin T, et al. Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. 2017 Mar 21;9(1):26. IF=7.071; pkt MNiSW=40

[D5] Liu P, Yuan B, Carvalho CM, Wuster A, Walter K, et al. An Organismal CNV Mutator Phenotype Restricted to Early Human Development. Cell. 2017 Feb 23;168(5):830-842.e7. IF=30.41; pkt MNiSW=50

[D6] Küry S, Besnard T, Ebstein F, Khan TN, Gambin T, et al. *De Novo* Disruption of the Proteasome Regulatory Subunit PSMD12 Causes a Syndromic Neurodevelopmental Disorder. Am J Hum Genet. 2017 Feb 2;100(2):352-363. IF=9.025; pkt MNiSW=45

[D7] Stray-Pedersen A, Sorte HS, Samarakoon P, Gambin T, Chinn IK, et al. Primary immunodeficiency diseases: Genomic approaches delineate heterogeneous Mendelian disorders. J Allergy Clin Immunol. 2017 Jan;139(1):232-245. IF=13.081; pkt MNiSW=50

[D8] Macias A, Gambin T, Szafranski P, Jhangiani SN, Kolasa A, et al. CAV3 mutation in a patient with transient hyperCKemia and myalgia. Neurol Neurochir Pol. 2016 Nov - Dec;50(6):468-473. IF=0.857 ; pkt MNiSW=15

[D9] Eldomery MK, Akdemir ZC, Vögtle FN, Charng WL, Mulica P, et al. MIPEP recessive variants cause a syndrome of left ventricular non-compaction, hypotonia, and infantile death. Genome Med. 2016 Nov 1;8(1):106. IF=7.071; pkt MNiSW=40

[D10] Sorte HS, Osnes LT, Fevang B, Aukrust P, Erichsen HC, et al. A potential founder variant in CARMIL2/RLTPR in three Norwegian families with warts, molluscum contagiosum, and T-cell dysfunction. Mol Genet Genomic Med. 2016 Nov;4(6):604-616. IF=2.979; pkt MNiSW=25

[D11] Prescott TE, Kulseth MA, Heimdal KR, Stadheim B, Hopp E, et al. Two male sibs with severe micrognathia and a missense variant in MED12. Eur J Med Genet. 2016 Aug;59(8):367-72. IF=2.137; pkt MNiSW=15

[D12] Charng WL, Karaca E, Coban Akdemir Z, Gambin T, Atik MM, et al. Exome sequencing in mostly consanguineous Arab families with neurologic disease provides a high potential molecular diagnosis rate. BMC Med Genomics. 2016 Jul 19;9(1):42. IF=2.198; pkt MNiSW=20

[D13] Gawlinski P, Posmyk R, Gambin T, Sielicka D, Chorazy M, et al. PEHO Syndrome May Represent Phenotypic Expansion at the Severe End of the Early-Onset Encephalopathies. Pediatr Neurol. 2016 Jul;60:83-7. IF=2.018; pkt MNiSW=25

[D14] Posey JE, Rosenfeld JA, James RA, Bainbridge M, Niu Z, et al. Molecular diagnostic experience of whole-exome sequencing in adult patients. Genet Med. 2016 Jul;18(7):678-85. IF=7.329 ; pkt MNiSW=40

[D15] Szafranski P, Gambin T, Dharmadhikari AV, Akdemir KC, Jhangiani SN, et al. Pathogenetics of alveolar capillary dysplasia with misalignment of pulmonary veins. Hum Genet. 2016 May;135(5):569-86. IF=4.637 ; pkt MNiSW=35

[D16] Karolak JA, Gambin T, Rydzanicz M, Szaflik JP, Polakowski P, et al. Evidence against ZNF469 being causative for keratoconus in Polish patients. Acta Ophthalmol. 2016 May;94(3):289-94. IF=3.157; pkt MNiSW=35

[D17] Lalani SR, Liu P, Rosenfeld JA, Watkin LB, Chiang T, et al. Recurrent Muscle Weakness with Rhabdomyolysis, Metabolic Crises, and Cardiac Arrhythmia Due to Bi-allelic TANGO2 Mutations. Am J Hum Genet. 2016 Feb 4;98(2):347-57. IF=9.025; pkt MNiSW=45

[D18] Bayram Y, Karaca E, Coban Akdemir Z, Yilmaz EO, Tayfun GA, et al. Molecular etiology of arthrogryposis in multiple families of mostly Turkish origin. J Clin Invest. 2016 Feb;126(2):762-78. IF=12.784; pkt MNiSW=50

[D19] Karolak JA, Gambin T, Pitarque JA, Molinari A, Jhangiani S, et al. Variants in SKP1, PROB1, and IL17B genes at keratoconus 5q311-q353 susceptibility locus identified by whole-exome sequencing. Eur J Hum Genet. 2016 Jan;25(1):73-78. IF=4.287; pkt MNiSW=35

[D20] Farlow JL, Robak LA, Hetrick K, Bowling K, Boerwinkle E, et al. Whole-Exome Sequencing in Familial Parkinson Disease. JAMA Neurol. 2016 Jan;73(1):68-75. IF=10.029; pkt MNiSW=45

[D21] Boone PM, Yuan B, Gu S, Ma Z, Gambin T, et al. Hutterite-type cataract maps to chromosome 6p2132-p2131, cosegregates with a homozygous mutation in LEMD2, and is associated with sudden cardiac death. Mol Genet Genomic Med. 2016 Jan;4(1):77-94. IF=2.979; pkt MNiSW=25

[D22] Yuan B, Liu P, Gupta A, Beck CR, Tejomurtula A, et al. Comparative Genomic Analyses of the Human NPHP1 Locus Reveal Complex Genomic Architecture and Its Regional Evolution in Primates. PLoS Genet. 2015 Dec;11(12):e1005686. IF=6.1; pkt MNiSW=45

[D23] Gambin M, Gambin T, Sharp C. Social cognition, psychopathological symptoms, and family functioning in a sample of inpatient adolescents using variable-centered and person-centered approaches. J Adolesc. 2015 Dec;45:31-43. IF=1.795; pkt MNiSW=30

[D24] Karaca E, Harel T, Pehlivan D, Jhangiani SN, Gambin T, et al. Genes that Affect Brain Structure and Function Identified by Rare Variant Analyses of Mendelian Neurologic Disease. Neuron. 2015 Nov 4;88(3):499-513. IF=14.024; pkt MNiSW=50

[D25] Karaca E, Yuregir OO, Bozdogan ST, Aslan H, Pehlivan D, et al. Rare variants in the notch signaling pathway describe a novel type of autosomal recessive Klippel-Feil syndrome. Am J Med Genet A. 2015 Nov;167A(11):2795-9. IF=2.259; pkt MNiSW=20

[D26] Bayram Y, Aydin H, Gambin T, Akdemir ZC, Atik MM, et al. Exome sequencing identifies a homozygous C5orf42 variant in a Turkish kindred with oral-facial-digital syndrome type VI. Am J Med Genet A. 2015 Sep;167A(9):2132-7. IF=2.259; pkt MNiSW=20

[D27] Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015 Aug 6;97(2):199-215. IF=9.025; pkt MNiSW=45

[D28] Bozdogan ST, Yuregir OO, Buyukkurt N, Aslan H, Ozdemir ZC, et al. Alpha-thalassemia mutations in adana province, southern Turkey: genotype-phenotype correlation. Indian J Hematol Blood Transfus. 2015 Jun;31(2):223-8. IF=0.403; pkt MNiSW=15

[D29] Watkin LB, Jessen B, Wiszniewski W, Vece TJ, Jan M, et al. COPA mutations impair ER-Golgi transport and cause hereditary autoimmune-mediated lung disease and arthritis. Nat Genet. 2015 Jun;47(6):654-60. IF=27.959; pkt MNiSW=50

[D30] Bayram Y, Gulsuner S, Guran T, Abaci A, Yesil G, et al. Homozygous loss-of-function mutations in SOHLH1 in patients with nonsyndromic hypergonadotropic hypogonadism. J Clin Endocrinol Metab. 2015 May;100(5):E808-14. IF=5.455; pkt MNiSW=40

[D31] White J, Mazzeu JF, Hoischen A, Jhangiani SN, Gambin T, et al. DVL1 frameshift mutations clustering in the penultimate exon cause autosomal-dominant Robinow syndrome. Am J Hum Genet. 2015 Apr 2;96(4):612-22. IF=9.025; pkt MNiSW=45

[D32] Beck TF, Campeau PM, Jhangiani SN, Gambin T, Li AH, et al. FBN1 contributing to familial congenital diaphragmatic hernia. Am J Med Genet A. 2015 Apr;167A(4):831-6. IF=2.259; pkt MNiSW=20

[D33] Beck CR, Carvalho CM, Banser L, Gambin T, Stubbolo D, et al. Complex genomic rearrangements at the PLP1 locus include triplication and quadruplication. PLoS Genet. 2015 Mar;11(3):e1005050. IF=6.1; pkt MNiSW=45

[D34] Riveiro-Álvarez R, Xie YA, López-Martínez MÁ, Gambin T, Pérez-Carro R, et al. New mutations in the RAB28 gene in 2 Spanish families with cone-rod dystrophy. JAMA Ophthalmol. 2015 Feb;133(2):133-9. IF=5.625; pkt MNiSW=45

[D35] Yuan B, Pehlivan D, Karaca E, Patel N, Charng WL, et al. Global transcriptional disturbances underlie Cornelia de Lange syndrome and related phenotypes. J Clin Invest. 2015 Feb;125(2):636-51. IF=12.784; pkt MNiSW=50

[D36] Karaca E, Buyukkaya R, Pehlivan D, Charng WL, Yaykasli KO, et al. Whole-exome sequencing identifies homozygous GPR161 mutation in a family with pituitary stalk interruption syndrome. J Clin Endocrinol Metab. 2015 Jan;100(1):E140-7. IF=5.455; pkt MNiSW=40

[D37] Collison FT, Xie YA, Gambin T, Jhangiani S, Muzny D, et al. Whole Exome Sequencing Identifies an Adult-Onset Case of Methylmalonic Aciduria and Homocystinuria Type C (cblC) with Non-Syndromic Bull's Eye Maculopathy. Ophthalmic Genet. 2015;36(3):270-5. IF=1.277; pkt MNiSW=20

[D38] Dharmadhikari AV, Gambin T, Szafranski P, Cao W, Probst FJ, et al. Molecular and clinical analyses of 16q241 duplications involving FOXF1 identify an evolutionarily unstable large minisatellite. BMC Med Genet. 2014 Dec 4;15:128. IF=2.198; pkt MNiSW=20

[D39] Xie YA, Lee W, Cai C, Gambin T, Nõupuu K, et al. New syndrome with retinitis pigmentosa is caused by nonsense mutations in retinol dehydrogenase RDH11. Hum Mol Genet. 2014 Nov 1;23(21):5774-80. IF=5.34; pkt MNiSW=40

[D40] Stray-Pedersen A, Jouanguy E, Crequer A, Bertuch AA, Brown BS, et al. Compound heterozygous CORO1A mutations in siblings with a mucocutaneous-immunodeficiency syndrome of epidermodysplasia verruciformis-HPV, molluscum contagiosum and granulomatous tuberculoid leprosy. J Clin Immunol. 2014 Oct;34(7):871-90. IF=3.253; pkt MNiSW=25

[D41] Yamamoto S, Jaiswal M, Charng WL, Gambin T, Karaca E, et al. A drosophila genetic resource of mutants to study mechanisms underlying human genetic diseases. Cell. 2014 Sep 25;159(1):200-14. IF=30.41; pkt MNiSW=50

[D42] Pehlivan D, Karaca E, Aydin H, Beck CR, Gambin T, et al. Whole-exome sequencing links TMCO1 defect syndrome with cerebro-facio-thoracic dysplasia. Eur J Hum Genet. 2014 Sep;22(9):1145-8. IF=4.287; pkt MNiSW=35

[D43] Bayram Y, Pehlivan D, Karaca E, Gambin T, Jhangiani SN, et al. Whole exome sequencing identifies three novel mutations in ANTXR1 in families with GAPO syndrome. Am J Med Genet A. 2014 Sep;164A(9):2328-34. IF=2.259; pkt MNiSW=20

[D44] Pham J, Shaw C, Pursley A, Hixson P, Sampath S, et al. Somatic mosaicism detected by exon-targeted, high-resolution aCGH in 10,362 consecutive cases. Eur J Hum Genet. 2014 Aug;22(8):969-78. IF=4.287; pkt MNiSW=35

[D45] Stray-Pedersen A, Backe PH, Sorte HS, Mørkrid L, Chokshi NY, et al. PGM3 mutations cause a congenital disorder of glycosylation with severe immunodeficiency and skeletal dysplasia. Am J Hum Genet. 2014 Jul 3;95(1):96-107. IF=9.025; pkt MNiSW=45

[D46] Karaca E, Weitzer S, Pehlivan D, Shiraishi H, Gogakos T, et al. Human CLP1 mutations alter tRNA biogenesis, affecting both peripheral and central nervous system function. Cell. 2014 Apr 24;157(3):636-50. IF=30.41; pkt MNiSW=50

[D47] Wangler MF, Gonzaga-Jauregui C, Gambin T, Penney S, Moss T, et al. Heterozygous *de novo* and inherited mutations in the smooth muscle actin (ACTG2) gene underlie megacystis-microcolon-intestinal hypoperistalsis syndrome. PLoS Genet. 2014 Mar;10(3):e1004258. IF=6.1; pkt MNiSW=45

[D48] Bartnik M, Nowakowska B, Derwińska K, Wiśniowiecka-Kowalnik B, Kędzior M, et al. Application of array comparative genomic hybridization in 256 patients with developmental delay or intellectual disability. J Appl Genet. 2014 Feb;55(1):125-44. IF=1.655; pkt MNiSW=20

[D49] Wiszniewska J, Bi W, Shaw C, Stankiewicz P, Kang SH, et al. Combined array CGH plus SNP genome analyses in a single assay for optimized clinical testing. Eur J Hum Genet. 2014 Jan;22(1):79-87. IF=4.287; pkt MNiSW=35

[D50] Wiśniowiecka-Kowalnik B, Kastory-Bronowska M, Bartnik M, Derwińska K, Dymczak-Domini W, et al. Application of custom-designed oligonucleotide array CGH in 145 patients with autistic spectrum disorders. Eur J Hum Genet. 2013 Jun;21(6):620-5. IF=4.287; pkt MNiSW=35

[D51] Bartnik M, Szczepanik E, Derwińska K, Wiśniowiecka-Kowalnik B, Gambin T, et al. Application of array comparative genomic hybridization in 102 patients with epilepsy and additional neurodevelopmental disorders. Am J Med Genet B Neuropsychiatr Genet. 2012 Oct;159B(7):760-71. IF=3.258; pkt MNiSW=30

[D52] Derwińska K, Bartnik M, Wiśniowiecka-Kowalnik B, Jagła M, Rudziński A, et al. Assessment of the role of copy-number variants in 150 patients with congenital heart defects. Med Wieku Rozwoj. 2012 Jul-Sep;16(3):175-82. Pkt MNiSW=7

Literatura dodatkowa

Aebersold, Ruedi, and Lars Malmstroem. 2013. "Faculty of 1000 Evaluation for An Integrated Map of Genetic Variation from 1,092 Human Genomes." *F1000 - Post-Publication Peer Review of the Biomedical Literature*. doi:10.3410/f.717971074.793469443.

Alekseyenko, Alexander V., and Christopher J. Lee. 2007. "Nested Containment List (NCList): A New Algorithm for Accelerating Interval Query of Genome Alignment and Interval Databases." *Bioinformatics* 23 (11): 1386–93.

Bailey, J. A., A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler. 2001. "Segmental Duplications: Organization and Impact within the Current Human Genome Project Assembly." *Genome Research* 11 (6): 1005–17.

Boone, Philip M., Carlos A. Bacino, Chad A. Shaw, Patricia A. Eng, Patricia M. Hixson, Amber N. Pursley, Sung-Hae L. Kang, et al. 2010. "Detection of Clinically Relevant Exonic Copy-Number Changes by Array CGH." *Human Mutation* 31 (12): 1326–42.

Cock, Peter J. A., Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. 2010. "The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants." *Nucleic Acids Research* 38 (6): 1767–71.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27 (15): 2156–58.

Eilers, P. H. C., and R. X. de Menezes. 2004. "Quantile Smoothing of Array CGH Data." *Bioinformatics* 21 (7): 1146–53.

Fromer, Menachem, Jennifer L. Moran, Kimberly Chambert, Eric Banks, Sarah E. Bergen, Douglas M. Ruderfer, Robert E. Handsaker, et al. 2012. "Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth." *American Journal of Human Genetics* 91 (4): 597–607.

Fromer, Menachem, and Shaun M. Purcell. 2014. "Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data." *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* 81 (April): 7.23.1–21.

Guo, Yan, Shilin Zhao, Brian D. Lehmann, Quanhu Sheng, Timothy M. Shaver, Thomas P. Stricker, Jennifer A. Pietenpol, and Yu Shyr. 2014. "Detection of Internal Exon Deletion with Exon Del." *BMC Bioinformatics* 15 (October): 332.

Jiang, Y., D. A. Oldridge, S. J. Diskin, and N. R. Zhang. 2015. "CODEX: A Normalization and Copy Number Variation Detection Method for Whole Exome Sequencing." *Nucleic Acids Research* 43 (6): e39–e39.

Karolchik, Donna, Angie S. Hinrichs, and W. James Kent. 2007. "The UCSC Genome Browser." In *Current Protocols in Bioinformatics*.

Krumm, Niklas, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, NHLBI Exome Sequencing Project, Aaron R. Quinlan, Deborah A. Nickerson, and Evan E. Eichler. 2012. "Copy Number Variation Detection and Genotyping from Exome Sequence Data." *Genome Research* 22 (8): 1525–32.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. "Circos: An Information Aesthetic for Comparative Genomics." *Genome Research* 19 (9): 1639–45.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. "Software for Computing and Annotating Genomic Ranges." *PLoS Computational Biology* 9 (8): e1003118.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic

Variation in 60,706 Humans." *Nature* 536 (7616): 285–91.

Ligt, Joep de, Philip M. Boone, Rolph Pfundt, Lisenka E. L. M. Vissers, Todd Richmond, Joel Geoghegan, Kathleen O'Moore, et al. 2013. "Detection of Clinically Relevant Copy Number Variants with Whole-Exome Sequencing." *Human Mutation* 34 (10): 1439–48.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Li, Jian, R. Alan Harris, Sau Wai Cheung, Cristian Coarfa, Mira Jeong, Margaret A. Goodell, Lisa D. White, et al. 2012. "Genomic Hypomethylation in the Human Germline Associates with Selective Structural Mutability in the Human Genome." *PLoS Genetics* 8 (5): e1002692.

Liu, Pengfei, Claudia M. B. Carvalho, P. J. Hastings, and James R. Lupski. 2012. "Mechanisms for Recurrent and Complex Human Genomic Rearrangements." *Current Opinion in Genetics & Development* 22 (3): 211–20.

Liu, Xiaoming, Chunlei Wu, Chang Li, and Eric Boerwinkle. 2016. "dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs." *Human Mutation* 37 (3): 235–41.

Lupski, J. R. 1998. "Genomic Disorders: Structural Features of the Genome Can Lead to DNA Rearrangements and Human Disease Traits." *Trends in Genetics: TIG* 14 (10): 417–22.

Morgenthaler, Stephan, and William G. Thilly. 2007. "A Strategy to Discover Genes That Carry Multi-Allelic or Mono-Allelic Risk for Common Diseases: A Cohort Allelic Sums Test (CAST)." *Mutation Research* 615 (1-2): 28–56.

O'Brien, Aidan R., Neil F. W. Saunders, Yi Guo, Fabian A. Buske, Rodney J. Scott, and Denis C. Bauer. 2015. "VariantSpark: Population Scale Clustering of Genotype Information." *BMC Genomics* 16 (December): 1052.

Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler. 2004. "Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data." *Biostatistics* 5 (4): 557–72.

Packer, Jonathan S., Evan K. Maxwell, Colm O'Dushlaine, Alexander E. Lopez, Frederick E. Dewey, Rostislav Chernomorsky, Aris Baras, John D. Overton, Lukas Habegger, and Jeffrey G. Reid. 2016. "CLAMMS: A Scalable Algorithm for Calling Common and Rare Copy Number Variants from Exome Sequencing Data." *Bioinformatics* 32 (1): 133–35.

Parsons, J. D. 1995. "Miropeats: Graphical DNA Sequence Comparisons." *Computer Applications in the Biosciences: CABIOS* 11 (6): 615–19.

Rentería, Miguel E., Adrian Cortes, and Sarah E. Medland. 2013. "Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis." In *Methods in Molecular Biology*, 193–213.

Santorico, Stephanie A., and Audrey E. Hendricks. 2016. "Progress in Methods for Rare Variant Association." *BMC Genetics* 17 (S2). doi:10.1186/s12863-015-0316-7.

Smith, Lacey A., Jessica Douglas, Alicia A. Braxton, and Kate Kramer. 2015. "Reporting Incidental Findings in Clinical Whole Exome Sequencing: Incorporation of the 2013 ACMG Recommendations into Current Practices of Genetic Counseling." *Journal of Genetic Counseling* 24 (4): 654–62.

Stankiewicz, Pawel, and James R. Lupski. 2002. "Genome Architecture, Rearrangements and Genomic Disorders." *Trends in Genetics: TIG* 18 (2): 74–82.

Van der Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. "From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline." In *Current Protocols in Bioinformatics*, 11.10.1–11.10.33.

Wu, Michael C., Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. 2011. "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test." *American Journal of Human Genetics* 89 (1): 82–93.

Zheng, Xiuwen, David Levine, Jess Shen, Stephanie M. Gogarten, Cathy Laurie, and Bruce S. Weir. 2012. "A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data." *Bioinformatics* 28 (24): 3326–28.