AUTOREFERAT

dr inż. Tomasz Stokowy

Centrum Analiz Biostatystycznych i Bioinformatycznych
Gdański Uniwersytet Medyczny
Gdańsk 2020

1. Imię i nazwisko:

Tomasz Stokowy

2. Posiadane dyplomy, stopnie naukowe – z podaniem nazwy, miejsca i roku ich uzyskania oraz tytułu rozprawy doktorskiej:

2009 magister inżynier; Macrocourse: Automation and Robotics, Electronics and Telecommunication, Computer Science; praca magisterska pt.: „Classification of DNA microarray data with random forests"; promotor: prof. dr hab. inż. Krzysztof Fujarewicz; studia w języku angielskim.

2013 doktor nauk technicznych; biocybernetyka i inżynieria biomedyczna; praca doktorska pt.: „Selection of miRNA isoform markers differentiating between follicular thyroid cancer and follicular thyroid adenoma from high-throughput sequencing data"; promotor: prof. dr hab. inż. Krzysztof Fujarewicz

3. Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych:

2009-2013 Studia doktoranckie, Wydział Automatyki, Informatyki i Elektroniki, Politechnika Śląska w Gliwicach

2009-2013 Studia doktoranckie, Zakład Medycyny Nuklearnej i Endokrynologii Onkologicznej, Narodowy Instytut Onkologii im. M. Skłodowskiej-Curie, Państwowy Instytut Badawczy Oddział w Gliwicach

2013-2016 Staż podoktorski, Department of Clinical Science, University of Bergen, Norway

2016-2017 Laboratory Associate, Gerstein Lab, Yale School of Medicine, CT, USA

Od 2016 Senior Engineer, Department of Clinical Science, University of Bergen, Norway

Od 2020 Specjalista naukowo techniczny, Centrum Analiz Biostatystycznych i Bioinformatycznych, Gdański Uniwersytet Medyczny

4. Osiągnięcie naukowe wynikające z art. 219 ust. 1 pkt. 2 Ustawy Prawo o szkolnictwie wyższym i nauce (Dz.U.2020.0.85 ze zm.):

Osiągnięcie naukowe wynikające z 219 ust. 1 pkt. 2 Ustawy Prawo o szkolnictwie wyższym i nauce stanowi cykl 5 powiązanych tematycznie publikacji dotyczących analizy i interpretacji rzadkich wariantów w danych pochodzących z głębokiego sekwencjonowania DNA. Prace te, publikowane w latach 2016-2019 są rezultatem współpracy z zespołami badawczymi w Polsce i za granicą.

Łączna wartość współczynnika oddziaływania IF prac składających się na osiągnięcie wynosi 27,841. Łączna liczba punktów MNiSW prac składających się na osiągnięcie wynosi 270 (według załącznika do komunikatu Ministra Nauki i Szkolnictwa Wyższego z dnia 31 lipca 2019 r.).

a) tytuł osiągnięcia naukowego:

Precyzyjna identyfikacja rzadkich wariantów genetycznych w danych pochodzących z sekwencjonowania DNA wysokiej przepustowości.

b) publikacje wchodzące w skład osiągnięcia naukowego:

4.1. **Stokowy T**, Garbulowski M, Fiskerstrand T, Holdhus R, Labun K, Sztromwasser P, Gilissen C, Hoischen A, Houge G, Petersen K, Jonassen I, Steen VM. RareVariantVis: new tool for visualization of causative variants in rare monogenic disorders using whole genome sequencing data. Bioinformatics. 2016 Oct 1;32(19):3018-20. doi: 10.1093/bioinformatics/btw359. PMID: 27288501. IF 7,307; MNiSW 45.

4.2. Ngcungcu T, Oti M, Sitek JC, Haukanes BI, Linghu B, Bruccoleri R, **Stokowy T**, Oakeley EJ, Yang F, Zhu J, Sultan M, Schalkwijk J, van Vlijmen-Willems IMJJ, von der Lippe C, Brunner HG, Ersland KM, Grayson W, Buechmann-Moller S, Sundnes O, Nirmala N, Morgan TM, van Bokhoven H, Steen VM, Hull PR, Szustakowski J, Staedtler F, Zhou H, Fiskerstrand T, Ramsay M. Duplicated Enhancer Region Increases Expression of CTSB and Segregates with Keratolytic Winter Erythema in South African and Norwegian Families. Am J Hum Genet. 2017 May 4;100(5):737-750. doi: 10.1016/j.ajhg.2017.03.012. PMID: 28457472. IF 8,855; MNiSW 45.

4.3. Supernat A, Vidarsson OV, Steen VM, **Stokowy T**. Comparison of three variant callers for human whole genome sequencing. Sci Rep. 2018 Dec 14;8(1):17851. doi: 10.1038/s41598-018-36177-7. PMID: 30552369. IF 4,011; MNiSW 40.

4.4. Bredrup C, **Stokowy T**, McGaughran J, Lee S, Sapkota D, Cristea I, Xu L, Tveit KS, Høvding G, Steen VM, Rødahl E, Bruland O, Houge G. A tyrosine kinase-activating variant Asn666Ser in PDGFRB causes a progeria-like condition in the severe end of Penttinen syndrome. Eur J Hum Genet. 2019 Apr;27(4):574-581. doi: 10.1038/s41431-018-0323-z. PMID: 30573803. IF 3,657; MNiSW 100.

4.5. **Stokowy T**, Polushina T, Sønderby IE, Karlsson R, Giddaluru S, Le Hellard S, Bergen SE, Sullivan PF, Andreassen OA, Djurovic S, Hultman CM, Steen VM. Genetic variation in 117 myelination-related genes in schizophrenia: Replication of association to lipid biosynthesis genes. Sci Rep. 2018 May 2;8(1):6915. doi: 10.1038/s41598-018-25280-4. PMID: 29720671. IF 4,011; MNiSW 40.

Oświadczenia habilitanta dotyczące wykonanych prac znajdują się w załączniku nr 3. Oświadczenia współautorów publikacji określające indywidualny wkład autorów w powstanie poszczególnych publikacji zamieszczono w załączniku nr 5.
W pracach zawarto wyniki uzyskane podczas realizacji trzyletniego podoktorskiego stażu badawczego Deep Sequencing in Biomedicine, finansowanego przez Trond Mohn Foundation (grant nr 807964). Projekt był realizowany w jednostce Genomics Core Facility, Department of Clinical Science, University of Bergen, Norway, której kierownikiem jest prof. Vidar Martin Steen.

c) omówienie celu naukowego ww. prac i osiągniętych wyników wraz z omówieniem ich ewentualnego wykorzystania

Wstęp

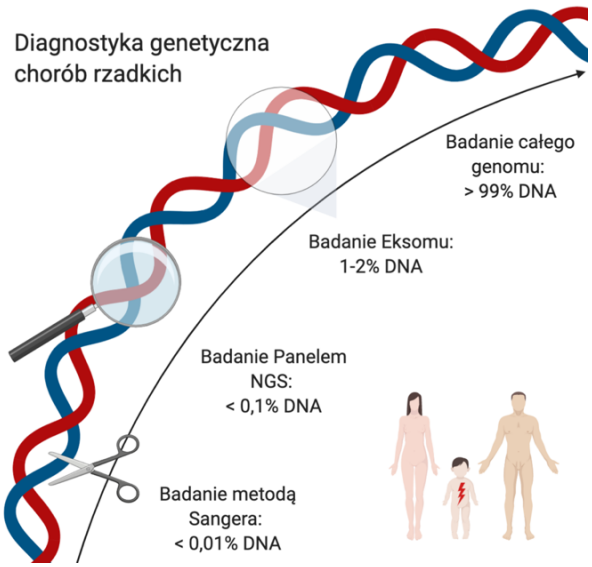
W Europie rzadkie choroby genetyczne (ang. rare diseases) są zdefiniowane jako dotykające mniej niż 1 na 2000 osób (Regulacja EC No 141/2000 on orphan medicinal products) [1]. Dotychczas opisanych zostało ponad 7000 chorób rzadkich. Około 5% ludzkiej populacji cierpi na jedną z tych chorób, co oznacza że w tylko w Europe jakiś rodzaj schorzenia genetycznego ma 30 milionów osób (https://www.orpha.net). Fenotypy chorób rzadkich są heterogenne i mają zróżnicowane przyczyny genetyczne. Choroby te dotykają głównie dzieci i są najczęściej zagrożeniem życia. Według bazy danych OMIM (https://www.omim.org) dla wielu chorób rzadkich podłoże genetyczne jest wciąż niepoznane.

Wyzwania pojawiające się w chorobach rzadkich wymagają najdokładniejszej możliwej diagnostyki – sekwencjonowania całego genomu (ang. whole genome sequencing, WGS, Rycina 1). Dzięki rozwojowi tej metody pacjenci z rzadkimi chorobami uzyskali jeden precyzyjny i szybki test, który pozwala na właściwą klasyfikację ich schorzenia. WGS pozwala dokładnie identyfikować różnego rodzaju warianty (od pojedynczego nukleotydu do dużych zmian strukturalnych). Dzięki temu metoda WGS stopniowo zastępuje kosztowne i pracochłonne analizy cytogenetyczne, MLPA (eng. Multiplex ligation-dependent probe amplification), analizy z wykorzystaniem mikromacierzy i techniki sekwencjonowania badające mniejsze spektrum genomowe [2].

Sekwencjonowanie wysokiej przepustowości (ang. Next Generation Sequencing, NGS) zrewolucjonizowało diagnostykę chorób, w których istotną rolę odgrywa czynnik genetyczny. Dzięki wprowadzeniu tej nowej technologii, w precyzyjny i szybki sposób można diagnozować rzadkie choroby genetyczne i nowotworowe. Badania NGS pozwalają na precyzyjną diagnostykę wykorzystującą najczęściej jedną z dostępnych opcji:

- badanie całego genomu (WGS), > 99% DNA
- badanie całego eksomu (WES), 1-2% DNA, wszystkie geny kodujące białka
- badanie panelem NGS, <0.1% DNA, oparte o wybrane spektrum od kilku do kilkuset genów

Dla porównania, przed wprowadzeniem technologii NGS w badaniach diagnostycznych, wykonywano najczęściej sekwencjonowanie Sangera, testujące 1 gen, co stanowi mniej niż 0.01% DNA (Rycina 1). Badania WES, panelowe i sekwencjonowaniem Sangera są tańsze niż WGS, lecz nie są tak precyzyjne i nie mają tak szerokiego spektrum diagnostycznego (WES nie pokrywa niekodującej części genomu). W literaturze można znaleźć przykłady potwierdzające, że WGS jest zdecydowanie dokładniejszą metodą niż WES [3,4] za cenę wyższego kosztu badania [5].



Rycina 1. Spektrum diagnostycznych badań DNA opartych o techniki sekwencjonowania. Źródło: www.genetyka.bio, wydanie 2/2020; autor Tomasz Stokowy

Diagnostyka personalizowana (precyzyjna) zakłada dobór leczenia rzadkich chorób genetycznych i nowotworowych w oparciu o unikatową informację o każdym pacjencie, w tym aberrację w sekwencji DNA, która leży u podłoża danej choroby. W badaniach DNA konieczne jest aby identyfikacja wariantów w danych NGS było wykonywana z najwyższą możliwą dokładnością. Zadanie to jest szczególnie trudne w wielkoskalowych danych z eksperymentów WGS i stanowi podstawę opisywanych badań. Wyniki Christiana Gilissena i współpracowników [3] oraz moja wizyta naukowa na uniwersytecie Radboud w Nijmegen zainspirowały moje badania nad metodą WGS, prowadzone w latach 2014-2020.

Cel badań

Celem prowadzonych badań przedstawionych w cyklu powiązanych tematycznie prac było stworzenie i wykorzystanie nowych metod analizy genomu do precyzyjnej identyfikacji wariantów DNA będących przyczyną chorób o podłożu genetycznym.

Wyniki

Praca 1

W roku 2014 gdy rozpoczynałem analizę próbek od pacjentów z chorobami rzadkimi sekwencjonowanymi metodą WGS dostępność narzędzi do analizy bioinformatycznej danych była ograniczona. Narzędzia były w większości dedykowane metodzie WES i nie nadawały się do danych WGS o wielkości ponad 50 GB danych na próbkę. W celu rozwiązania tego problemu stworzyłem, zaimplementowałem i opublikowałem narzędzie RareVariantVis [6]. Metoda ta i jej publikacja w czasopiśmie Bioinformatics jest pierwszym elementem osiągnięcia naukowego (4.1). Narzędzie służy do filtrowania, adnotacji i wizualizacji rzadkich wariantów genetycznych i regionów utraty heterozygotyczności (ang. LOH, Loss of Heterozygosity). Pozwala na analizę wariantów genetycznych wywołanych takimi narzędziami jak GATK [7], SpeedSeq/Freebayes [8] czy DeepVariant [9]. Pozwala też użytkownikowi określać i dobierać własne parametry filtrowania wariantów. Rzadkie, niesynonimiczne kodujące warianty są zwracane w formie tabeli i wizualizowane na odpowiednich chromosomach. Unikatowość tego podejścia polega na łączeniu automatycznego filtrowania wariantów z możliwością wizualizacji i manualnej weryfikacji wyników w kontekście całych chromosomów. W niektórych przypadkach pozwala to dodatkowo identyfikować skomplikowane warianty strukturalne, których nie udało się zaobserwować innymi metodami. Pakiet jest dostępny w repozytorium Bioconductor, pod adresem:
https://www.bioconductor.org/packages/release/bioc/html/RareVariantVis.html.

Mój wkład w powstanie tej pracy polegał na zaproponowaniu hipotezy badawczej, zaprojektowaniu i programowaniu biblioteki RareVariantVis w języku R, zgłoszeniu biblioteki do Bioconductor, pozyskaniu danych testowych, testowaniu i przygotowaniu wyników do artykułu oraz napisaniu manuskryptu.

Praca 2

Zastosowanie narzędzia RareVariantVis pozwoliło na odkrycie nieznanych dotąd przyczyn kilku rzadkich chorób genetycznych. Pierwszą z nich był keratolityczny rumień zimowy (eng. Keratolytic Winter Erythema OMIM: https://www.omim.org/entry/148370, KWE). Choroba ta charakteryzuje się powtarzającym się sezonowo łuszczeniem skóry, szczególnie w okresie zimnej i wilgotnej pogody (Rycina 2). Dzięki zastosowaniu sekwencjonowania całego genomu (WGS) i zaimplementowanych przeze mnie metod analizy udało się odkryć przyczynę choroby: duplikację wzmacniacza (eng. Enhancer) w regionie genu *CTSB*. Gen ten koduje proteazę cysteinową odpowiedzialną za homeostazę keratynocytów, podobnie jak inne katapsyny z rodziny związane z chorobami skóry (*CTSA i CTSC*). Odkrycie genetycznej przyczyny KWE pozwoliło na rozpoczęcie diagnostyki genetycznej KWE i lepsze zrozumienie funkcji genu *CTSB* [10].

Istotnym aspektem odkrycia jest to, że wariant będący przyczyną choroby znajduje się w regionie niekodującym. Należy podkreślić, że wariant nie został wykryty wcześniejszymi badaniami metodami WES i sekwencjonowaniem Sangera. Proces wykrycia wariantu był utrudniony ze względu na fakt że jest to duplikacja tandemowa – jeden z najtrudniejszych do informatycznego wywołania wariantów genetycznych [11]. Wykorzystanie narzędzia RareVariantVis i metody CNVnator [12] było podstawą odkrycia opublikowanego w czasopiśmie American Journal of Human Genetics (4.2) [10]. Praca ta dała podstawy dalszych badań funkcjonalnych nad aspektami genetycznymi chorób skóry, prowadzonymi na Uniwersytecie w Bergen i Witwatersrand University w Johannesburgu, RPA.

Rycina 2. Fenotyp Keratolytic Winter Erythema powodowany przez duplikację wzmacniacza genu CTSB. Ngcungcu T., (…), Stokowy T. et al. [10]

Mój wkład w powstanie tej pracy polegał na odkryciu wariantu we wzmacniaczu genu CTSB u pacjentów norweskich, analizie danych RNA i przygotowaniu wyników analizy funkcjonalnej, przygotowaniu manuskryptu do publikacji.

Praca 3

Podczas pracy z germinalnymi wariantami genetycznymi zauważyłem, że bardzo trudno zdecydować jakich narzędzi analizy użyć żeby uzyskać maksymalną dokładność wywoływania wariantów genetycznych. Publikacje naukowe nie wskazywały w jasny i wiarygodny sposób najlepszych metod analizy i często realizowane były we współpracy z firmami promującymi własne rozwiązanie. Komercyjne analizy często nie ukazują w pełni wad metody promowanej przez firmę, dlatego niezależna akademicka ewaluacja pomaga w niezależnej ocenie rozwiązania technologicznego. Aby rozwiązać ten problem przeprowadziłem i opublikowałem test porównawczy (eng. benchmark) oceniający w niezależny sposób jakość, precyzję, czułość i specyficzność najnowszych metod wywoływania wariantów genetycznych.

W trzeciej pracy będącej elementem mojego dzieła naukowego wykazuję, że DeepVariant jest najdokładniejszą obecnie dostępną metodą wywoływania wariantów pojedynczego nukleotydu (SNV) i krótkich insercji/delecji (4.3) [13]. W pracy tej realizowanej wspólnie z dr Anną Supernat opieram się o dane z wystandaryzowanej próbki NA12878 zsekwencjonowanej w Genomics Core Facility na Uniwersytecie w Bergen. Próbka ta jest wykorzystywana także w innych badaniach tworzących standardy, na przykład Genome in a Bottle (GIAB, opracowane przez National Institute of Standards and Technology, USA) i testach wariantów FDA (Food and Drug Administration, Federal Agency USA).

Metoda DeepVariant [9] jako pierwsza wywołuje warianty wykorzystując sztuczną inteligencję, a dokładniej technikę głębokiego uczenia maszynowego (deep learning). Zastosowanie tej metody pozwoliło na dokładniejsze wywołanie wariantów niż metody uznawane dotychczas za złoty standard [7]. Obecnie wyniki mojej pracy (4.3) zostały zaktualizowane i rozszerzone we współpracy z zespołem z Uniwersytetu w Oslo [14]. W aktualnych badaniach uwzględniliśmy serwer Dragen firmy Illumina, który pozwala na precyzyjne mapowanie i wywołanie wariantów dla całego genomu w mniej niż 30 minut.

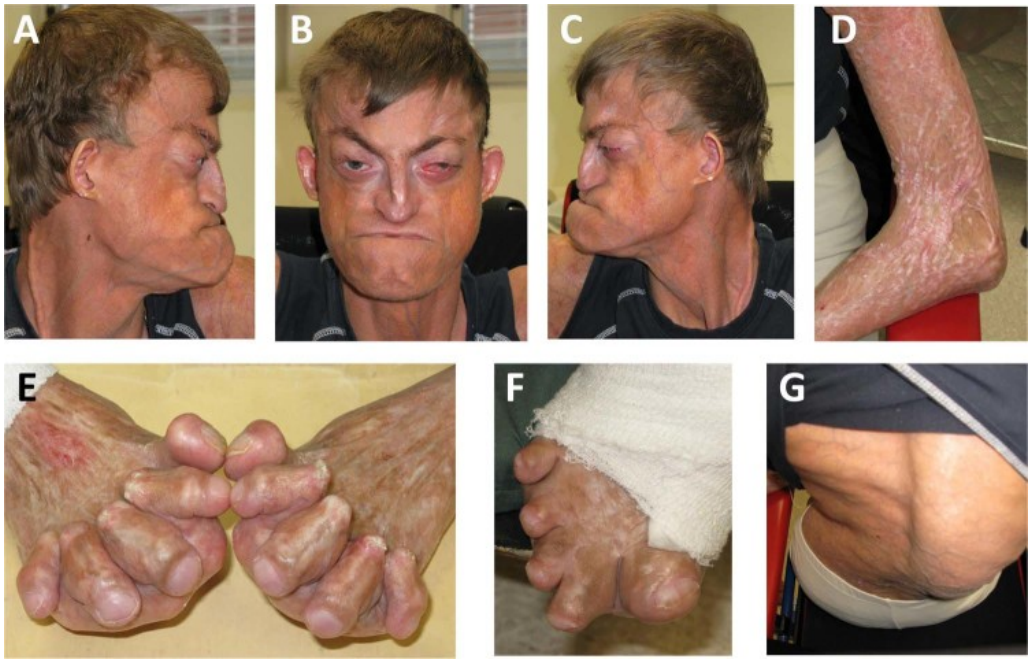
Mój wkład w powstanie tej pracy polegał na zaprojektowaniu badania, zebraniu zespołu badawczego, analizie danych z głębokiego sekwencjonowania, zebraniu wyników do manuskryptu, nadzorowaniu projektu, napisaniu manuskryptu i korekcie jego finalnej wersji.

Praca 4

Zastosowanie samodzielnie zaimplementowanych metod (4.1) oraz właściwa identyfikacja najlepszych narzędzi do wywoływania wariantów (4.3) pozwoliły na odkrycie istotnych wariantów genetycznych w syndromie Penttinena (4.4)(OMIM #601812) [15]. Dzięki zastosowaniu WGS udało się znaleźć wariant będący przyczyną choroby, który nie został wcześniej wykryty metodą WES ze względu na niskie pokrycie regionu.
Wariant typu *de novo* odkryty dzięki wykonanej przeze mnie analizie WGS znajduje się w genie *PDGFRB* (platelet derived growth receptor beta), c.1997A>G p.(Asn666Ser).

Dokładniejsza lokalizacja wariantu wskazała na domenę kinazy tyrozynowej *PDGFRB*, związanej bezpośrednio z syndromami przedwczesnego starzenia (Penttinen-type i Penttinen-like). Odkryty ultra rzadki wariant (identyczny u pacjenta norweskiego i australijskiego) daje wyjątkowo silny fenotyp choroby i charakteryzuje się poważnym zniszczeniem tkanek łącznych i charakterystycznym zniekształceniem kończyn (Ryciny 3 i 4).

Rycina 3. Fenotyp powodowany przez warianty w receptorze o aktywności kinazy tyrozynowej PDGFRB u pacjentki norweskiej. Bredrup C., Stokowy T. et al. [15]



Rycina 4. Fenotyp powodowany przez warianty w receptorze o aktywności kinazy tyrozynowej PDGFRB u pacjenta australijskiego. Bredrup C., Stokowy T. et al.

Badania funkcjonalne wykonane w ramach projektu pozwoliły na zidentyfikowanie potencjalnej opcji terapeutycznej dla pacjentów z tą chorobą – Imatynibu. W pracy opisującej wyniki badania połączono nowatorską diagnostykę WGS z doborem terapii, więc wpisuje się ona w spektrum badań medycyny personalizowanej. Mój wkład w powstanie tej pracy polegał na zaprojektowaniu analizy danych, pozyskaniu danych, analizie danych z całych genomów, odkryciu wariantu będącego przyczyną choroby, przygotowaniu manuskryptu do publikacji.

Praca 5

Badania nad wariantami rzadkimi oprócz chorób dziedziczonych według praw Mendla mają swoje zastosowanie w onkologii i chorobach populacyjnych. W piątej publikacji wchodzącej w skład osiągnięcia naukowego opisuję rolę rzadkich i częstych wariantów genetycznych w schizofrenii. W pracy (4.5)[16] analizowałem warianty związane z biosyntezą kwasów tłuszczowych, odgrywających rolę w procesie mielinizacji u pacjentów ze schizofrenią.
Schizofrenia jest chorobą o silnym podłożu dziedzicznym, jednak powody tego dziedziczenia nie są dobrze poznane [17,18]. Występuje element „brakującego dziedziczenia" (ang. missing heritability), czyli różnicy pomiędzy procentowym rzeczywistym dziedziczeniem choroby i dziedziczeniem wytłumaczonym naukowo.

W pracy podkreślam rolę wariantów w genach *SREBF1* i *SREBF2* i rzadkich wariantów w genie *LRP1* weryfikując wyniki wcześniej uzyskane przez współpracowników z mojej grupy badawczej [19]. Warianty w tych genach osiągnęły poziom nominalnej znamienności w porównaniu chorych i zdrowych kontroli ($p < 0.05$), jednak nie przetrwały korekty na wielokrotne testowanie. Podsumowując, przeprowadzone analizy wykazały że warianty w genach związanych z mielinizacją nie są głównym czynnikiem ryzyka dla schizofrenii.

Mój wkład w powstanie tej pracy polegał na opracowaniu hipotezy badawczej, pozyskaniu danych, wykonaniu analizy wyników, komunikacji ze współpracownikami w Szwecji i przygotowaniu manuskryptu.

Najważniejsze wnioski pochodzące z cyklu publikacji

- Przyczyną keratolitycznego rumienia zimowego (OMIM #148370) jest duplikacja wzmacniacza genu *CTSB* odkryta w regionie niekodującym genomu.
- Wariant c.1997A>G p.(Asn666Ser) w genie *PDGFRB* powoduje poważny fenotyp syndromu Penttinena. Imatynib jest potencjalną opcją terapeutyczną dla pacjentów z tym syndromem.
- Sekwencjonowanie całego genomu pozwala na dokładne wywoływanie małych wariantów genetycznych (>98% dla wariantów pojedynczego nukleotydu i >94% dla wariantów indel). Metoda DeepVariant wykorzystująca sztuczną inteligencję jest dokładniejsza niż inne dotychczas wykorzystywane standardy analityczne.
- Metoda RareVariantVis pozwala na odkrywanie nowych wariantów będących przyczyną chorób genetycznych w danych z sekwencjonowania całych genomów.
- Rzadkie warianty genetyczne związane z mielinizacją nie są głównym czynnikiem ryzyka odpowiedzialnym za dziedziczenie schizofrenii.

Praktyczne zastosowanie wyników badań

Wyniki opisanych badań pozwalają na wdrożenie diagnostyki genomowej dla pacjentów z chorobami rzadkimi. Przykładem takiego wdrożenia jest uruchomienie w Poznaniu startupu naukowego MNM Diagnostics, który w swojej ofercie ma badanie chorób rzadkich metodą WGS. W roku 2018 zostałem współzałożycielem MNM Diagnostics, wnosząc wiedzę techniczną dotyczącą badań genomowych. Powstanie startupu umożliwiło uruchomienie nowatorskich metod diagnostyki chorób rzadkich w Polsce (https://mnm.bio/pl/produkty-i-uslugi/diagnostyka-chorob-rzadkich/). Oprócz chorób dziedzicznych firma wykonuje obecnie diagnostykę onkologiczną, profilowanie genomowe, rozwija biobankowanie i wspiera walkę z wirusem COVID-19. Z perspektywy praktycznego zastosowania wyników badań istotna jest także edukacja społeczeństwa, która jest prowadzona przez portal popularnonaukowy Fakty i mity genetyki (https://genetyka.bio).

Z globalnego punktu widzenia głównym osiągnięciem jest odkrycie przyczyn Keratolytic Winter Erythema i wariantów powodujących poważny fenotyp syndromu Penttinena. Przez publikację wyników prac w międzynarodowych czasopismach i bazie OMIM pacjenci z tymi chorobami będą mogli uzyskać właściwą diagnozę w oparciu o WGS i inne metody diagnostyki molekularnej. Odkrycie jest istotne także dla mniejszych ośrodków diagnostycznych, które preferują tańsze rozwiązania diagnostyczne niż WGS (na przykład panele skoncentrowane na konkretnym regionie genomowym).

Wprowadzenie nowych i ocena wcześniej opublikowanych metod analizy WGS pozwoliły na kolejny krok w rozwoju diagnostyki genomowej. Moje prace i uzyskane wyniki były elementem tego kroku w kierunku lepszego poznania genomu człowieka. Należy jednak zaznaczyć, że technologie sekwencjonowania rozwijają się bardzo szybko, a ceny analiz

spadają. Efektem tego jest rozwój technologii i pojawianie się coraz nowszych i lepszych metod analitycznych. Jestem przekonany że kolejne lata pozwolą na stworzenie i przygotowanie jeszcze dokładniejszych metod diagnostycznych, szczególnie w zakresie analizy wariantów strukturalnych i wykorzystania długich odczytów.

Wykorzystana metodologia

W pracach niniejszego cyklu publikacji zastosowano szerokie spektrum metod analitycznych i samodzielnie zaimplementowanego oprogramowania.
W analizach statystycznych, filtrowaniu, adnotacji wariantów, i wizualizacji danych wykorzystywałem samodzielnie zaimplementowane narzędzia w środowisku R/Bioconductor (prace 4.1, 4.2, 4.5). Narzędzia udostępniłem w Bioconductor, są dostępne w formie open access:
https://www.bioconductor.org/packages/release/bioc/html/RareVariantVis.html.
Do mapowania odczytów z sekwencjonowania, wywoływania wariantów i adnotacji wykorzystywałem skrypty Unix shell, zaimplementowane w systemie SAFE na Uniwersytecie w Bergen: https://it.uib.no/ithjelp/images/b/b2/SAFE_E_-_For_decision_makers.pdf
System SAFE pozwalał na bezpieczne przetworzenie i przechowywanie danych wrażliwych pochodzących od pacjentów (prace 4.2, 4.4, 4.5) oraz efektywną ocenę jakości danych WGS (praca 4.3). Ze względu na prawo w Norwegii i ograniczenia udostępniania danych personalnych analizy wykonywane były w środowisku bez dostępu do Internetu.
Do wywoływania wariantów germinalnych wykorzystywałem głównie metodę DeepVariant opierającą się o sztuczną inteligencję i uczenie maszynowe. Do przyspieszenia tych analiz wykorzystywałem akceleratory graficzne Nvidia (opisane szczegółowo w pracy 4.3).
Narzędzia w wielu przypadkach implementowane i wykorzystywane były w technologii kontenerów Docker, co pozwala na swobodne przenoszenie do innych systemów eksperymentalnych i diagnostycznych.

W mojej pracy wykorzystałem szereg publicznie dostępnych narzędzi, które pozwoliły na odkrycie przyczyn rzadkich chorób genetycznych w danych z sekwencjonowania całych genomów:

- speedseq [8]
- DeepVariant [9]
- GATK [7]
- hap.py – Haplotype Comparison Tools [20]
- VariantAnnotation [21]
- Variant Effect Predictor [22]
- CNVnator [12]
- ANNOVAR [23]

Bibliografia

1. Regulation (EC) No 141/2000 of the European Parliament and o... - EUR-Lex Available online: https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX%3A32000R0141 (accessed on Sep 13, 2020).
2. Lindstrand, A.; Eisfeldt, J.; Pettersson, M.; Carvalho, C.M.B.; Kvarnung, M.; Grigelioniene, G.; Anderlid, B.-M.; Bjerin, O.; Gustavsson, P.; Hammarsjö, A.; et al. From cytogenetics to cytogenomics: whole-genome sequencing as a first-line test comprehensively captures the diverse spectrum of disease-causing genetic variation underlying intellectual disability. *Genome Med.* **2019**, *11*, 68, doi:10.1186/s13073-019-0675-1.
3. Gilissen, C.; Hehir-Kwa, J.Y.; Thung, D.T.; van de Vorst, M.; van Bon, B.W.M.; Willemsen, M.H.; Kwint, M.; Janssen, I.M.; Hoischen, A.; Schenck, A.; et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **2014**, *511*, 344–347, doi:10.1038/nature13394.
4. Lelieveld, S.H.; Spielmann, M.; Mundlos, S.; Veltman, J.A.; Gilissen, C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum. Mutat.* **2015**, *36*, 815–822, doi:10.1002/humu.22813.
5. Schwarze, K.; Buchanan, J.; Taylor, J.C.; Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* **2018**, *20*, 1122–1130, doi:10.1038/gim.2017.247.
6. Stokowy, T.; Garbulowski, M.; Fiskerstrand, T.; Holdhus, R.; Labun, K.; Sztromwasser, P.; Gilissen, C.; Hoischen, A.; Houge, G.; Petersen, K.; et al. RareVariantVis: new tool for visualization of causative variants in rare monogenic disorders using whole genome sequencing data. *Bioinforma. Oxf. Engl.* **2016**, *32*, 3018–3020, doi:10.1093/bioinformatics/btw359.
7. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303, doi:10.1101/gr.107524.110.
8. Chiang, C.; Layer, R.M.; Faust, G.G.; Lindberg, M.R.; Rose, D.B.; Garrison, E.P.; Marth, G.T.; Quinlan, A.R.; Hall, I.M. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **2015**, *12*, 966–968, doi:10.1038/nmeth.3505.
9. Poplin, R.; Chang, P.-C.; Alexander, D.; Schwartz, S.; Colthurst, T.; Ku, A.; Newburger, D.; Dijamco, J.; Nguyen, N.; Afshar, P.T.; et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **2018**, *36*, 983–987, doi:10.1038/nbt.4235.
10. Ngcungcu, T.; Oti, M.; Sitek, J.C.; Haukanes, B.I.; Linghu, B.; Bruccoleri, R.; Stokowy, T.; Oakeley, E.J.; Yang, F.; Zhu, J.; et al. Duplicated Enhancer Region Increases Expression of CTSB and Segregates with Keratolytic Winter Erythema in South African and Norwegian Families. *Am. J. Hum. Genet.* **2017**, *100*, 737–750, doi:10.1016/j.ajhg.2017.03.012.
11. Kosugi, S.; Momozawa, Y.; Liu, X.; Terao, C.; Kubo, M.; Kamatani, Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **2019**, *20*, doi:10.1186/s13059-019-1720-5.
12. Abyzov, A.; Urban, A.E.; Snyder, M.; Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **2011**, *21*, 974–984, doi:10.1101/gr.114876.110.
13. Supernat, A.; Vidarsson, O.V.; Steen, V.M.; Stokowy, T. Comparison of three variant callers for human whole genome sequencing. *Sci. Rep.* **2018**, *8*, doi:10.1038/s41598-018-36177-7.
14. Zhao, S.; Agafonov, O.; Azab, A.; Stokowy, T.; Hovig, E. Accuracy and efficiency of germline variant calling pipelines for human genome data. *Sci. Rep.* **2020**, *10*, 20222, doi:10.1038/s41598-020-77218-4.

15. Bredrup, C.; Stokowy, T.; McGaughran, J.; Lee, S.; Sapkota, D.; Cristea, I.; Xu, L.; Tveit, K.S.; Høvding, G.; Steen, V.M.; et al. A tyrosine kinase-activating variant Asn666Ser in PDGFRB causes a progeria-like condition in the severe end of Penttinen syndrome. *Eur. J. Hum. Genet. EJHG* **2019**, *27*, 574–581, doi:10.1038/s41431-018-0323-z.
16. Stokowy, T.; Polushina, T.; Sønderby, I.E.; Karlsson, R.; Giddaluru, S.; Le Hellard, S.; Bergen, S.E.; Sullivan, P.F.; Andreassen, O.A.; Djurovic, S.; et al. Genetic variation in 117 myelination-related genes in schizophrenia: Replication of association to lipid biosynthesis genes. *Sci. Rep.* **2018**, *8*, 6915, doi:10.1038/s41598-018-25280-4.
17. Ripke, S.; Neale, B.M.; Corvin, A.; Walters, J.T.R.; Farh, K.-H.; Holmans, P.A.; Lee, P.; Bulik-Sullivan, B.; Collier, D.A.; Huang, H.; et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **2014**, *511*, 421–427, doi:10.1038/nature13595.
18. Li, Z.; Chen, J.; Yu, H.; He, L.; Xu, Y.; Zhang, D.; Yi, Q.; Li, C.; Li, X.; Shen, J.; et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **2017**, *49*, 1576–1583, doi:10.1038/ng.3973.
19. Le Hellard, S.; Mühleisen, T.W.; Djurovic, S.; Fernø, J.; Ouriaghi, Z.; Mattheisen, M.; Vasilescu, C.; Raeder, M.B.; Hansen, T.; Strohmaier, J.; et al. Polymorphisms in SREBF1 and SREBF2, two antipsychotic-activated transcription factors controlling cellular lipogenesis, are associated with schizophrenia in German and Scandinavian samples. *Mol. Psychiatry* **2010**, *15*, 463–472, doi:10.1038/mp.2008.110.
20. *Illumina/hap.py*; Illumina, 2020;
21. Obenchain, V.; Lawrence, M.; Carey, V.; Gogarten, S.; Shannon, P.; Morgan, M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinforma. Oxf. Engl.* **2014**, *30*, 2076–2078, doi:10.1093/bioinformatics/btu168.
22. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biol.* **2016**, *17*, 122, doi:10.1186/s13059-016-0974-4.
23. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164, doi:10.1093/nar/gkq603.

5. Omówienie pozostałych osiągnięć naukowo badawczych

H-Index Web of Science / Scopus:	14 / 15
Liczba cytowań bez autocytowań Web of Science / Scopus:	489 / 539
Sumaryczny Impact Factor po uzyskaniu stopnia doktora:	148,807

Moje badania nad rzadkimi wariantami pozwoliły na identyfikację istotnych wariantów genetycznych w kilku badaniach związanych z rzadkimi chorobami (załącznik 3, wykaz osiągnięć naukowych). Wyniki opisane zostały między innymi w moich pracach opisujących wykorzystanie sekwencjonowania egzomów w pierwotnych niedoborach odporności (Arts et al., Genome Medicine 2019) oraz przyczyn syndromu progerii związanego z *LEMD2* (Marbach et al., AJHG 2019). Prace zostały przygotowane we współpracy z wiodącymi europejskimi ośrodkami zajmującymi się tematyką chorób rzadkich.

Oprócz powyższych prac o wysokim współczynniku oddziaływania (impact factor) byłem zaangażowany w szereg prac z dziedziny genetyki onkologicznej. Prace te powstały głównie w związku ze współpracą z Gdańskim Uniwersytetem Medycznym, ale także z Radboud University Nijmegen (Neveling et al., Clin Chem 2017). W dużej mierze prace te dotyczą raka piersi i jajnika, gdzie poznanie komponentu genetycznego jest istotnym elementem doboru terapii.

Wyniki mojej pracy przedstawiałem także podczas licznych konferencji i wizyt naukowych:

- Genome Informatics 2015, Cold Spring Harbor Laboratory, NY, USA (prezentacja ustna)
- The Biology of Genomes 2017, Cold Spring Harbor Laboratory, NY, USA (prezentacja plakatowa)
- American Society of Human Genetics 2018, San Diego, CA, USA (prezentacja plakatowa)
- Intelligent Systems for Molecular Biology 2019, Basel, Szwajcaria (prezentacja plakatowa)
- Yale University, CT, USA (prezentacja dla grupy badawczej profesora Marka Gerstein, z którą byłem związany w roli laboratory associate w latach 2016-2017)
- Stanford University, CA, USA (prezentacja dla grupy badawczej profesora Matta van de Rijn podczas wizyty naukowej w Kalifornii w roku 2018)
- Uniwersytety w Oslo, Trondheim oraz Tromsø, w ramach cyklu spotkań National Consortium for Sequencing and Personalized Medicine w Norwegii.

W wyniku mojej współpracy z dr Anną Supernat jestem także współautorem jednego zgłoszenia patentowego do Urzędu Patentowego Rzeczpospolitej Polskiej zatytułowanego „Sposób analizy i klasyfikacji materiału biologicznego w wykrywaniu choroby nowotworowej”. Zgłoszenie oznaczono zostało numerem: P.435990

6. Projekty badawcze i badawczo - rozwojowe

Od początku mojej działalności naukowej zaangażowany byłem w szereg projektów badawczych i badawczo-rozwojowych, głównie w roli wykonawcy:

- "Deep sequencing in biomedicine" – Trond Mohn Foundation (BFS2016-genom); wykonawca
- "National consortium for sequencing and personalized medicine" – Research Council of Norway (245979/F50); wykonawca
- "Analysis of interaction between cancer cells and Tumor Educated Platelets" – NCN (2018/31/D/NZ5/01263); wykonawca
- "RareVariantVis2: whole human genome analysis suite" Meltzer Fund (ID 16363); kierownik
- "Novel gene biomarkers of spermatogenesis – potential for spermatogenesis assessment and treatment monitoring" – NCN (2012/05/N/NZ5/00893); wykonawca
- „Opracowanie molekularnych testów wspomagających wykrywanie wczesnego raka płuca" – Gdański Uniwersytet Medyczny – MOLTEST2013; wykonawca
- "Molecular Genomics, Transcruptomics and Bioinformatics in Cancer" – Fundacja Nauki Polskiej (MPD 2009/5); wykonawca
- „Wpływ czynnika transkrypcyjnego HSF1 na transformację nowotworową indukowaną przez estrogen" – NCN (2015/17/B/NZ3/03760); wykonawca
- "Novel causative genetic variants in azoospermia: whole genome analysis and functional in vitro studies" NCN (2017/26/D/NZ5/00789); wykonawca
- „Profilaktyka i leczenie chorób cywilizacyjnych STRATEGMED" – NCBR (STRATEGMED2/267398/4/NCBR/2015); wykonawca
- „Inicjatywa doskonałości – Uczelnia Badawcza", Gdański Uniwersytet Medyczny (IDUB 2020); wykonawca
- „CanCell Cancer – Zapobieganie rozwojowi chorób nowotworowych poprzez edukację zdrowotną" – Norway Grants 2009-2014, program Bilateral Cooperation Fund, Development and better adaptation of health care to demographic and epidemiological trends; partner

7. Przebieg pracy naukowej przed uzyskaniem stopnia doktora

Przed rozpoczęciem pracy naukowej studiowałem na kierunku Makrokierunek (Automatyka i robotyka, elektronika i telekomunikacja, informatyka) na Politechnice Śląskiej. Studia ukończyłem w specjalności Information processing for control broniąc dyplom magistra inżyniera i pracę zatytułowaną „Classification of DNA microarray data with random forests".

Po studiach magisterskich rozpocząłem studia doktoranckie na Politechnice Śląskiej w Zakładzie Inżynierii Systemów Instytutu Automatyki. Równolegle otrzymałem stypendium Fundacji Nauki Polskiej i Studium Medycyny Molekularnej Warszawskiego Uniwersytetu Medycznego. Dzięki stypendium rozpocząłem pracę w Narodowym Instytucie Onkologii, Oddział w Gliwicach, gdzie badałem molekularne cechy raka tarczycy. Stypendium dawało możliwość stażu zagranicznego, który odbyłem w latach 2012-2013 na Uniwersytecie w Lipsku, w Niemczech.

Studia doktoranckie pozwoliły mi zdobyć doświadczenie w zakresie analizy danych z głębokiego sekwencjonowania DNA i RNA Illumina i jego zastosowania w genetyce onkologicznej. Brałem udział w kursach i konferencjach naukowych, między innymi w kursie Integrative analysis of genome scale data – Cold Spring Harbor Laboratory, NY, USA (czerwiec 2010). Kurs pozwolił na dalszy rozwój moich zainteresowań badawczych w tym zakresie. Dzięki zastosowaniu poznanych technik obliczeniowych uzyskałem stypendium EMBL-EBI w celu prezentacji moich wyników na kongresie ISMB SCS 2012 w Los Angeles, CA, USA w roku 2012. Na kongresie zaprezentowałem pracę "RNA-Seq reveals usefulness of small RNA isoforms in thyroid tumors diagnosis".

Podczas moich studiów doktoranckich uzyskałem nieocenione wsparcie od opiekunów i mentorów: prof. Krzysztofa Fujarewicza, promotora mojej pracy magisterskiej i doktorskiej, dr hab. Michała Jarząba, promotora pomocniczego mojej pracy doktorskiej, prof. Andrzeja Świerniaka, prof. Barbary Jarząb, dr inż. Michała Świerniaka, prof. Ralfa Paschke i dr Markusa Eszlingera.

Moja praca nad molekularnymi markerami raka tarczycy zaowocowała kilkunastoma pełno tekstowymi publikacjami z tematyki raka tarczycy, z których większość przygotowaliśmy wspólnie z dr inż. Bartoszem Wojtasiem. Moją pracę doktorską zatytułowaną „Selection of miRNA isoform markers differentiating between follicular thyroid cancer and follicular thyroid adenoma from high-throughput sequencing data" obroniłem we wrześniu 2013. Dzięki dobrej współpracy naszego zespołu wspólne badania nad molekularnymi markerami raka tarczycy trwają do dziś.

8. Przebieg pracy naukowej po uzyskaniu stopnia doktora

8.1 Staż podoktorski

Bezpośrednio po uzyskaniu stopnia doktora zdecydowałem się na odbycie trzyletniego stażu podoktorskiego na Wydziale Medycyny Uniwersytetu w Bergen, w Norwegii. Staż był w całości finansowany przez Trond Mohn Foundation i zatytułowany „Deep sequencing in biomedicine". W ramach stażu pracowałem w jednostce Department of Clinical Science, zlokalizowanej w szpitalu uniwersyteckim w Bergen – Haukeland University Hospital, Department of Medical Genetics. Kierownikiem zespołu badawczego i mojego stażu podoktorskiego był prof. Vidar Martin Steen. Projekt realizowany był w ścisłej współpracy z prof. Gunnarem Houge, kierownikiem Department of Medical Genetics w szpitalu uniwersyteckim w Bergen oraz dr Torunn Fiskerstrand, zmarłą w roku 2019.

W trakcie stażu podoktorskiego odbyłem dwie wizyty naukowe: Radboud University Medical Center Nijmegen, Holandia (2014, w grupie prof. Alexa Hoischen) oraz Yale School of Medicine, New Haven, CT, USA (2016-2017, jako laboratory associate w grupie prof. Marka Gersteina). Wizyty te pozwoliły na rozwój moich zainteresowań badawczych związanych z analizą całego genomu (WGS) oraz rozwinęły umiejętności analityczne.

Podczas stażu podoktorskiego pracowałem nad analizami całego genomu człowieka, szczególnie w kontekście analizy rzadkich wariantów genetycznych. Prace te miały szerokie spektrum: rozwój metod analitycznych, ocenę dostępnych rozwiązań, wykorzystanie nowych narzędzi bioinformatycznych, odkrywanie nowych wariantów genetycznych będących przyczyną chorób i analizy funkcjonalne uzyskanych wyników. Wynikiem prowadzonych prac były publikacje naukowe wchodzące w skład osiągnięcia naukowego (4.1 – 4.5).

8.2 Stanowisko Senior Engineer

Po zakończeniu stażu podoktorskiego dostałem propozycję stałej pracy na stanowisku Senior Engineer w Genomics Core Facility na Uniwersytecie w Bergen. Jednostka ta zatrudnia 4 osoby i świadczy usługi sekwencjonowania, analiz bioinformatycznych i wsparcia projektów naukowych na Uniwersytecie w Bergen. Co roku realizuje ponad 50 projektów badawczo-rozwojowych. Core Facility jest częścią sieci NorSeq - The Norwegian Consortium for Sequencing and Personalized Medicine, w której wspólnie z partnerami z Oslo, Trondheim i Tromso realizujemy projekty z zakresu medycyny precyzyjnej w Norwegii.

W ramach Genomics Core Facility projektuję i implementuję rozwiązania informatyczne dla diagnostyki medycznej, głównie w zakresie chorób rzadkich i onkologii. Opracowane metody mają charakter przenośny, dzięki wykorzystaniu technologii kontenerów (np. Docker). To rozwiązanie pozwala na sprawne wdrażanie metod analitycznych do innych jednostek badawczych i diagnostycznych w kraju i za granicą.

Od roku 2019 pełnię także rolę konsultanta Centrum Analiz Biostatystycznych i Bioinformatycznych GUMed. Wspieram aktywnie rozwój projektu „Inicjatywa doskonałości – Uczelnia Badawcza" na Gdańskim Uniwersytecie Medycznym oraz zadania Zakładu Onkologii Translacyjnej GUMed w zakresie rozwoju technologii analizy płynnych biopsji.

Uczestniczę w pracach komisji oceniających wnioski grantowe i rekrutujących pracowników na stanowiska w instytucjach publicznych. Od roku 2019 jestem ekspertem Agencji Badań Medycznych finansującej niekomercyjne badania kliniczne w Polsce. Recenzowałem wnioski grantowe dla Research Council Faroe Islands oraz aplikacje do Studium Medycyny Molekularnej Warszawskiego Uniwersytetu Medycznego. W ramach pracy na Uniwersytecie w Bergen brałem udział w komisjach rekrutacyjnych na stanowiska Researcher, Engineer i Postdoctoral Fellow, w roli przewodniczącego i członka komisji.

9. Działalność dydaktyczna

Równolegle z działalnością naukową prowadziłem zajęcia dydaktyczne dla studentów i doktorantów.

- Politechnika Śląska w Gliwicach, studia magisterskie i inżynierskie, na kierunku Automatyka i Robotyka/Makrokierunek: Komputerowo wspomagane podejmowanie decyzji, Sieci Neuronowe, Metody sztucznej inteligencji, Optymalizacja i podejmowanie decyzji, Systemy biotechniczne, Sztuczna inteligencja i sieci neuronowe
- Uniwersytet w Bergen, studia magisterskie: Human Molecular Genetics
- Uniwersytet Adama Mickiewicza w Poznaniu, studia magisterskie: Wysokoprzepustowe technologie sekwencjonowania i ich wykorzystanie w badaniach biomedycznych, Konstruowanie biomedycznych baz danych
- Warszawski Uniwersytet Medyczny, Studium Medycyny Molekularnej – studium doktoranckie
- Gdański Uniwersytet Medyczny – szkoła letnia Biotechnology Summer School 2019

Tomasz Stokowy