Metody precyzyjnej lokalizacji i pomiaru zjawisk w genomach i transkryptomach za pomocą analizy danych z mikromacierzy i sekwencerów nowej generacji. Autoreferat.

Michał Jerzy Okoniewski

27 czerwca 2015

Spis treści

1 Dane osobowe i przebieg zatrudnienia w jednostkach naukowych

Imię i nazwisko: Michał Jerzy Okoniewski

Dyplomy i stopnie naukowe:
2002 - doktor nauk technicznych, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych 1999 - magister inżynier, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych

Przebieg zatrudnienia:

1999-2002 Politechnika Warszawska, Instytut Informatyki, doktorant

2002-2003 Politechnika Warszawska, Instytut Informatyki, adiunkt

2003-2004 Uniwersytet w Antwerpii, Wydział Matematyki. Staż po-doktorski

2005-2007 Paterson Institute for Cancer Research, Manchester. Staż po-doktorski

2012-2013 Universitatspital Zurich, Ekspert genomiki i integracji danych

2008-czerwiec 2014 Functional Genomics Center Zurich, Ekspert transkryptomiki

lipiec 2014 - obecnie Research Informatics, Scientific IT Services, ETH (Politechnika Federalna w Zurichu), Ekspert genomiki obliczeniowej

2 Osiągnięcie naukowe

Wskazanie osiągnięcia* wynikającego z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. nr 65, poz. 595 ze zm.)

2.1 Tytuł osiągnięcia naukowego

"Metody precyzyjnej lokalizacji i pomiaru zjawisk w genomach i transkryptomach za pomocą analizy danych z mikromacierzy i sekwencerów nowej generacji."

2.2 Wykaz publikacji stanowiących osiągniecie naukowe

Spis publikacji tworzących jednotematyczny cykl:

(* - oznacza wspólne "pierwsze" autorstwo)

[P1] **Okoniewski MJ.**, Miller C., Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations, BMC Bioinformatics, 7:276, 2006

IF 3.02, cytowań wg. Google Scholar: 111, WoS: 60

[P4] Yates T, **Okoniewski MJ.**, Miller CJ., X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. Nucleic Acids Res. 2008 Jan;36(Database issue):D780-6.

IF 8.27, cytowań wg. Google Scholar: 58, WoS: 35

[P5] **Okoniewski MJ.**, Miller CJ. Comprehensive analysis of Affymetrix exon arrays using BioConductor. PLoS Comput Biol. 2008 Feb;4(2):e6.

IF 4.86, cytowań wg. Google Scholar: 38, WoS: 25

[P7] Leśniewska A, **Okoniewski MJ.**, rnaSeqMap, a Bioconductor package for RNA sequencing exploration, BMC Bioinformatics 2011, 12:200

IF 3.02, cytowań wg. Google Scholar: 12, WoS: 6

[P11] Wiewiórka MS., Messina A., Pacholewska, A., Maffioletti S., Gawrysiak P., **Okoniewski MJ.** SparkSeq: fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision. Bioinformatics, 2014, 30 (18): 2652-2653

IF 4.61, cytowań wg. Google Scholar: 5, WoS: 0

2.3 Publikacje powiązane

Publikacje związane z głównym osiągnięciem naukowym, jednak z przyczyn formalnych (trudności z zebraniem wszystkich oświadczeń współautorów) nie zaliczone bezpośrednio do osiągnięcia.

[P2] **Okoniewski MJ.**, Hey Y., Pepper S., Miller C.: High correspondence between Affymetrix exon and standard expression arrays. Biotechniques, Volume 42, Number 2, 2007

IF 2.66, cytowań wg. Google Scholar: 58, WoS: 39

[P3] **Okoniewski MJ.**., Yates T., Dibben S., Miller C., An annotation infrastructure for the analysis and interpretation of Affymetrix exon array data, Genome Biology 2007, 8:R79

IF 10.3, cytowań wg. Google Scholar: 49, WoS: 32

[P6] Bitton D*, **Okoniewski MJ.***, Connolly Y, Miller C. Exon level integration of proteomics and microarray data.BMC Bioinformatics. 2008 Feb 25;9:118

IF 3.02, cytowań wg. Google Scholar: 28, WoS: 17

[P8] **Okoniewski MJ.**, Leśniewska A, Szabelska A, Zyprych-Walczak J, Ryan M,Wachtel M ,Morzy T, Schäffer, B, Schlapbach R, Preferred analysis methods for single genomic regions in RNA sequencing revealed by processing the shape of coverage, Nucleic Acids Research, 2011

IF 8.81, cytowań wg. Google Scholar: 8, WoS: 1

[P9] **Okoniewski MJ.***, Meienberg J*, Patrignani A, Szabelska A, Matyas G, Schlapbach R, Precise breakpoint localization of large genomic deletions using PacBio and Illumina next-generation sequencers. Biotechniques, 2013 Feb;54(2):98-100. doi: 10.2144/000113992

IF 2.66, cytowań wg. Google Scholar: 4, WoS: 0

[P10] Anders S, McCarthy DJ, Chen Y, **Okoniewski MJ**, Smyth GK, Huber W, Robinson MD, Count-based differential expression analysis of RNA sequencing data using R and Bioconductor, Nature Protocols 2013 Sep;8(9):1765-86.

IF 7.96, cytowań wg. Google Scholar: 99, WoS: 43

2.4 Recenzowane pakiety oprogramowania naukowego wchodzące w skład osiągnięcia naukowego

[S1] Biblioteka **exonmap** w repozytorium Bioconductor [1]. Pierwsza wersja w 2007. W wersji obecnej jako "annmap - Genome annotation and visualisation package pertaining to Affymetrix arrays and NGS analysis" [2].

[S2] Biblioteka **rnaSeqMap** w repozytorium Bioconductor: "rnaSeqMap - rnaSeq secondary analyses" [3]. Pierwsza wersja w 2010.

[S3] Biblioteka **ampliQueso** w repozytorium Bioconductor: "ampliQueso - Analysis of amplicon enrichment panels" [4]. Pierwsza wersja w 2013.

[S4] **SparkSeq**, pakiet otwartego oprogramowania naukowego [5]

[1] http://www.bioconductor.org/packages/2.0/bioc/html/exonmap.html
[2] http://bioconductor.org/packages/release/bioc/html/annmap.html
[3] http://bioconductor.org/packages/release/bioc/html/rnaSeqMap.html
[4] http://bioconductor.org/packages/release/bioc/html/ampliQueso.html
[5] https://bitbucket.org/mwiewiorka/sparkseq

2.5 Przyznane granty na prace badawcze wchodzące w skład osiągnięcia naukowego

[G1] SCIEX 09.025 Grant szwajcarski programu "Scientific Exchange Program CH-NMS", 2009. Tytuł: "SeqMap - Software pipeline for RNA sequencing", 30'000 CHF na półroczny pobyt dr Anny Leśniewskiej z Politechniki Poznańskiej w Functional Genomics Center Zurich. Praca w Zurichu stała się podstawą doktoratu p. Leśniewskiej. Projekt we współpracy z prof. Tadeuszem Morzym - Politechnika Poznańska.

[G2] SCIEX 11.182 Grant szwajcarski programu "Scientific Exchange Program CH-NMS", 2011. Tytuł: "ShapeRNASeq - statistical extraction of biological features from RNA sequencing data", 60'000 CHF na półroczny pobyt dr Alicji Szabelskiej z Uniwersytetu Przyrodniczego w Poznaniu w Functional Genomics Center Zurich. Praca w Zurichu stała się podstawą doktoratu p. Szabelskiej, którego byłem promotorem pomocniczym. Projekt we współpracy z promotorem głównym, prof. Idzim Siatkowskim - Uniwersytet Przyrodniczy w Poznaniu.

[G3] SCIEX 12.289 Grant szwajcarski programu "Scientific Exchange Program CH-NMS", 2012. Tytuł: "CLANG – ClinicaL ANalysis in Genomics - a Dedicated Application for Translational Biomedical Research", 60'000 CHF na półroczny pobyt mgr Marka Wiewiórki z Politechniki Warszawskiej w Szpitalu Uniwersyteckim w Zurichu. Projekt we współpracy z prof. Piotrem Gawrysiakiem - Politechnika Warszawska.

[G4] Microsoft Azure Reseach Award, 2014. Tytuł: "Towards an interactive secondary analysis of RNA sequencing data service in Widows Azure cloud with Apache Spark framework". 40'000 USD do wykorzystania na zasoby obliczeniowe chmury Microsoft Azure. Współautor projektu - mgr Marek Wiewiórka, Politechnika Warszawska.

[G5] Grant OPUS Narodowego Centrum Nauki, PerM-Cloud, Algorytmy i metody przetwarzania dużych zbiorów danych genomicznych w środowiskach chmur obliczeniowych na potrzeby personalizowanej medycyny. Politechnika Warszawska, Kierownik Projektu. 499'138 zł

3 Opis osiągnięcia naukowego

3.1 Precyzyjna analiza ekspresji genów z użyciem mikromacierzy 3'-IVT Affymetrix

Analiza danych z genomu stała się praktycznie możliwa od roku 1977, kiedy zastosowano po raz pierwszy metodę Sangera [1]. Stała się ona pod-

stawą przełomowych odkryć w biologii molekularnej, z którym jednym z ważniejszych jest poznanie w 2001 całego ludzkiego genomu w [2]. Oprócz sekwencjonowania Sangera, jedną z podstawowych technologii umożliwiających rozpoznawanie DNA i RNA w próbce stały się mikromacierze, których początki datuje się na początek lat 1980-tych [3]. Technologia mikromacierzy rozwijała się dynamicznie w ciągu następnych lat, poprzez mikromacierze z dwoma barwnikami fluorescencyjnymi [4] do mikromacierzy oligonukleotydowych [5], najpowszechniej używanych do dziś. Mikromacierze firmy Affymetrix, wykonane techniką fotolitografii, pozwalały na jednoczesny pomiar wielu tysięcy krótkich, 25-nukleotydowych sekwencji. Za pomocą mikromacierzy Affymetrix można mierzyć ekspresję RNA poznając poziom znacznika fluorescencyjnego pozostającego na mikromacierzy po procesie hybrydyzacji RNA w próbce do oligonukleotydowej sondy. Sondy, których sekwencje odpowiadały za pomiar określonego fragmentu genu nazywano "probeset" (zbiór sond). Wartość ekspresji dla zbioru sond wyznaczano za pomocą metod statystycznych uśredniających pomiary sond, takich jak MAS zawarte w oprogramowaniu Affymetrix, czy RMA [6]. W technologii mikromacierzy Affymemetrix 3'-IVT, większość zbiorów sond mierzyła ekspresję genów od strony 3'. Istniały jednak alternatywne zbiory sond, mierzące inne fragmenty genów. Znajomość ich ekspresji różnicowej pozwalała na bardzo ograniczoną dodatkową możliwość określenia występowania alternatywnego splicingu genu.

W technologii Affymetrix, w podobny sposób poprzez różnicę hybrydyzacji sondy PM (perfect match) i MM (mismatch, ze środkowym nukleotydem zmienionym na komplementarny) można było wyznaczać obecność w DNA ważnych polimorfizmów jednonukleotydowych (ang. SNP - single nucleotide polymporphism).

Technologie firmy Affymetrix dla RNA i DNA zapoczątkowały możliwość wyznaczania ekspresji RNA i wariantów DNA dla większości znanych genów równocześnie i rozróżniających regiony mierzone przez mikromacierz z dokładnością do pojedynczych nukleotydów.

Technologia ta była w powszechnym zastosowaniu w biologii molekularnej w latach 2000-2010, jednakże naukowcy wykorzystujący ją do znajdowania ekspresji RNA często utożsamiali uśredniony wynik pomiaru zbioru sond jako "ekspresję genu", co prowadziło do różnego rodzaju artefaktów obliczeniowych dla różnych grup pomiarów zbiorów sond. Jedną z bardzo popularnych metod prezentacji danych była mapa ciepła (ang. heatmap) otrzymana przez jednoczesne grupowanie hierarchiczne pomiarów ekspresji z próbek biologicznych w kolumnach i z pojedynczych zbiorów sond w rzędach. Miarą odległości używaną do grupowania dwu wektorów pomiarów jest najczęściej korelacja R Pearsona. Mapę ciepła wykonuje się najczęściej dla podzbioru wyników mikromacierzy, np. dla zbiorów sond z największą wariancją. Grupowanie zbiorów sond i próbek biologicznych ma odzwierciedlać współregulację genów mierzonych przez uporządkowane w grupy zbiory

sond. Z kolei ich wartości dla określonych prób biologicznych mają wiązać ich cechy z ekspresją genów.

W roku 2005, gdy rozpocząłem pracę w Paterson Institute of Cancer Research w Manchesterze, w grupie dra Crispina Millera. Postanowił on sprawdzić swoją hipotezę mówiącą, że mapy ciepła nie zawsze pokazują jedynie współregulację genów i polecił mi zbadanie możliwych artefaktualnych współzależności pomiędzy zbiorami próbek w mikromacierzach Affymetrix. Podstawą do rozpoczęcia badań była baza danych ADAPT [7] zawierająca sondy mikromacierzy Affymetrix oraz ich mapowanie do sekwencji znanych transkryptów (izoform) genów z baz RefSeq i ENSEMBL. Praca **[P1]** opisuje analizy struktury sekwencji mikromacierzy w kontekście bazy ADAPT i eksperymentu pomiaru ekspresji RNA z linii komórkowej raka piersi MCF7 porównywanej z linią komórkową MCF10A. W wyniku badań okazało się, że w schemacie projektowym wszystkich mikromacierzy 3'-IVT istnieją sondy mierzące sygnał (poziom ekspresji) wielu transkryptów jednocześnie. Zjawisko takie nazywa się hybrydyzacją krzyżową (ew. kroshybrydyzacją) w mikromacierzach. Sygnał nakłada się, co powoduje zmianę uśrednionej wartości ekspresji zbiorów sond. Zmiana ta nie wpływa najczęściej znacząco na poziom mierzonego sygnału. Jednak w wypadku zbiorów sond posiadających część wspólną sond, powoduje zwiększoną korelację ich uśrednionego sygnału. Posiadanie części wspólnej może wynikać z tego, że sondy zaprojektowane zostały w sekwencjach domen białkowych lub też w sekwencjach powtarzalnych na genomie. W pracy wykazałem, że efektem może być mylna interpretacja wyników na poziomie procesów biologicznych. Postulowanym rozwiązaniem było sprawdzanie zbiorów sond przed interpretacją biologiczną na obecność hybrydyzacji krzyżowej za pomocą grafów zbudowanych z danych baz takich jak ADAPT - inaczej niż np. w rozwiązaniu z pracy [8] gdzie algorytm odrzuca sondy z kroshybrydyzacją.

Praca **[P1]** wywołała oddźwięk w środowisku badaczy posługujących się mikromacierzami Affymetrix, była wielokrotnie cytowana, doczekała się nawet krytycznej analizy w pracy naukowców pochodzących z Rosji [9]. Praca o podobnym temacie, cytująca **[P1]** została opublikowana przez zespół z Gliwic [10].

3.2 Analiza ekspresji RNA z dokładnością do eksonów z użyciem mikromacierzy Affymetrix Exon 1.0 ST

Podczas prac nad **[P1]**, firma Affymerix wypuściła na rynek kolejną i praktycznie ostatnią generację mikromacierzy - mikromacierze eksonowe 1.0 ST. Dzięki większej gęstości sond na mikromacierzy, pozwalała na jednoczesny pomiar 6,5 miliona oligonukleotydowych sond. Dzięki nowszemu protokołowi laboratoryjnemu przygotowania próbek biologicznych stało się również możliwe precyzyjniejsze mierzenie ekspresji sondami zaprojektowanymi na całej długości genu, nie tylko w 3'. Sondy pogrupowane są w zbiory

po 4 (na macierzach 3'-IVT było to 11 czy 16). Każdy zbiór sond odpowiadał w procesie projektowania jednemu regionowi genomowemu uznanemu za ekson. Dane o położeniu eksonów na genomie otrzymano przez integrację danych z kilkunastu baz danych anotacji genomowej, wśród których wciąż najistotniejsze były RefSeq i Ensembl.

Analiza danych pochodzących z mikromacierzy eksonowych za pomocą oprogramowania dostarczonego przez Affymetrix nie była kompatybilna z metodami analiz mikromacierzy poprzednich generacji. Affymetrix zakładał użycie innego formatu danych i innego sposobu rozumienia wyników. Ekspresja sumaryzowana była dla "klastrów transkryptów", które w wielu przypadkach były praktycznie regionami określającymi geny. Aby przetestować użyteczność mikromacierzy eksonowych, wykonałem eksperyment obliczeniowy porównujący dane pochodzące z tych samych próbek biologicznych (linie komórkowe MCF7 i MCF10, trzy replikaty dla każdej z nich) zmierzonych za pomocą mikromacierzy HGU133plus2 i eksonowej Human 1.0 ST. Wyniki zostały opisane w pracy **[P2]**.

Dla trzech różnych metod połączenia genomowych regionów docelowych w obu mikromacierzach, znaleziona została ilość zbiorów sond z HGU133plus2 mających swoje odpowiedniki w macierzy eksonowej (do 80%), zaś spośród zbiorów sond pokazujących wykrywalną ekspresję ("detected probesets") 96% zachowywało różnicę zlogarytmowanych (log2) zmian ekspresji ("fold change") mniejszą niż 1. Wyniki pracy **[P2]** potwierdziły porównywalność odczytów między mikromacierzami obu generacji. Praca była cytowana w wielu artykułach opisujących użycie mikromacierzy eksonowych Affymetrix jako dowód ich użyteczności. Praca **[P2]** była wysłana do recenzji w *Biotechniques* jako najwcześniejszy artykuł o macierzach eksonowych, jednak długi proces edytorski spowodował, że pierwszym opublikowanym 2 miesiące wcześniej artykułem była praca naukowców z samej firmy Affymetrix [11]. Praca ta udowadniała przydatność mikromacierzy eksonowych do rozpoznawania wariantów splicingowych genów i definiowała miarę indeksu splicingowego (SI) określonego na każdym z mierzonych eksonów.

Kolejnym etapem prac nad wykorzystaniem pełni możliwości mikromacierzy eksonowych było zaprojektowanie systemu analizy danych X:MAP-exonmap. Wykorzystano w nim zmodyfikowaną bazę danych ENSEMBL (dla genomu człowieka, myszy i szczura). Oryginalne bazy ENSEMBL w formacie MySQL tworzone są we współpracy EBI/EMBL i Wellcome Trust Sanger Institute w Hinxton/Cambridge [6]. Baza została uzupełniona o tabele relacyjne opisujące mapowanie oligonukleotydówych sond mikromacierzy do genomu referencyjnego (podobnie jak w [7]) oraz tabele grupującą sondy w zbiory sond zgodnie z oryginalną anotacją projektową Affymetrix. Tak utworzona baza została uzupełniona o procedury wbudowane w SQL pozwalające na znajdowanie sond odpowiadającym eksonom, transkryptom i

[6] http://www.ensembl.org/downloads.html

genom oraz mapowanie odwrotne: znajdowanie sond odpowiadającym regionom genomowym.

Pierwszym interfejsem do analizy danych jaki utworzyłem wraz z Crispinem Millerem była biblioteka exonmap [**S1**], napisana w języku R, opublikowana po recenzji w repozytorium Bioconductor [7]. Pozwalała ona na łączenie wyników ekspresji mikromacierzy eksonowych z trójwarstwową anotacją (ekson-transkrypt-gen) z bazy X:MAP. Komunikacja pomiędzy bazą X:MAP a funkcjami w R odbywa się z użyciem funkcji z biblioteki RMySQL, która na naszą prośbę została zmodyfikową przez nas o możliwość wywoływania procedur wbudowanych MySQL. API w jezyku R zawarte w bibliotece exonmap pozwalało również na odfiltrowanie zbiorów sond zawierających sondy hybrydyzujące krzyżowo, zgodnie z definicją z pracy [**P1**]. Dzięki temu powstało narzędzie programistyczne umożliwiające zaawansowane analizy fragmentów genów i znajdowanie różnych form eskpresji RNA ujawniającej alternatywny splicing genów - zarówno na poziomie pojedynczych genów jak i całości genomu (np. analiza działania genów będących czynnikami splicingowymi).

Całość oprogramowania do analiz danych z mikromacierzy eksonowych zawarta w bibliotece exonmap i bazie X:MAP wraz z podstawami teoretycznymi została opisana w pracy [**P3**]. Szczegółowe zagadnienia budowy bazy X:MAP oraz prawdopodobnie pierwszą na świecie przeglądarkę (browser) genomu z użyciem Google Maps API, zostały opublikowane w pracy [**P4**] w dodatku poświęconym bazom danych czasopisma Nucleic Acids Research. Razem z Crispinem Millerem wygłosiłem czterogodzinny tutorial na konferencji ISMB (Intelligent Systems for Molecular Biology) w roku 2007. Podsumowaniem tego wykładu była praca [**P5**] opisująca głównie aspekty biblioteki exonmap istotne dla jej użytkowników: bioinformatyków-progamistów i biologów molekularnych oraz lekarzy badających ekspresję genów z dokładnością do eksonów i wariantów splicingowych. System X:MAP-exonmap został od 2007 roku dwukrotnie poddany re-engineeringowi przez zespół dra Millera, obecnie zarówno przeglądarka genomu jak i biblioteka w repozytorium Bioconductor nazywają się annmap [8]

3.3 Połączenie precyzyjnych danych opisujących transkrypcję RNA z precyzyjnymi odczytami proteomowymi

Metody analizy z biblioteki exonmap [**S1**] zostały wykorzystane do porównania ekspresji różnicowej transkryptów RNA z ekspresją różnicową ich produktów proteinowych, mierzonych za pomocą spektrometrii masowej. Zostało to opisane w pracy [**P6**]. Układ eksperymentalny obejmował po trzy próbki z linii komórkowych MCF7 i MCF10A, w których zmierzono ekspresjonowane RNA za pomocą mikromacierzy eksonowej i produkty białko-

[7] http://www.bioconductor.org/
[8] http://annmap.cruk.manchester.ac.uk/

we metodą iTRAQ. Analizy danych transkryptomowych wykonywane były przeze mnie, proteomowych przez Danny'ego Bittona. Oba rodzaje analiz łączone były ze sobą poprzez regiony genowe odpowiadające zbiorom sond i produktom proteinowym. Przetestowaliśmy cztery strategie łączenia danych transkryptomowych i proteomowych porównując podobieństwo (porównywalność) ekspresji różnicowej za pomocą korelacji Pearsona dla regionów, które można połączyć. W pracy **[P6]** wykazaliśmy, że najlepsza porównywalność jest przy zastosowaniu najmniejszej granularności (największej precyzji) regionów genomowych: zbiorów sond dla eksonów na mikromacierzy i poziomu ekspresji peptydów po stronie proteinowej. Poziom korelacji w tym przypadku wyniósł R=0,808, co było wówczas najwyższym wynikiem porównywalności transkryptomowo-proteomowej opisanym w pracy naukowej. Korelacja ta nie może wynosić 1 z powodu modyfikacji post-transkrypcyjnych i niedoskonałości oraz różnic metod pomiarowych i analizy danych.

3.4 Analiza danych z nukleotydową precyzją dla sekwencerów nowej generacji

Pracę w Functional Genomics Center Zurich (FGCZ) rozpocząłem w roku 2008. FGCZ jest laboratorium pomiarów genomowych i spektrometrii masowej dla Uniwersytetu Zurychskiego i ETH Zurich. W ramach tej misji zakupiło stopniowo wszystkie platformy sekwencerów nowej generacji dostępne na rynku od 2008 do 2013: Roche/454, SOLiD, Illumina HiSeq/MiSeq, Ion Torrent/Proton, a takze sekwencer trzeciej generacji, sekwencjonujący pojedyncze molekuły bez etapu klonowania - firmy Pacific Biosystems.

Od początku pracy w FGCZ stosowałem tam wyniki swojej pracy nad mikromacierzami Affymetrix i precyzyjną lokalizacją regionów ekspresji RNA i wariantów splicingowych. Wynikiem były projekty i artykuły naukowe we współpracy z różnymi laboratoriami biologicznymi i medycznymi z Zurichu, na potrzeby których wykonywałem analizy statystyczne danych z mikromacierzy i analizy funkcji biologicznej genów o ekspresji różnicującej.

Samodzielną pracę nad zagadnieniami sekwencjonowania nowej generacji rozpocząłem w 2010 po otrzymaniu grantu **[G1]** przeznaczonego na ukierunkowanie pracy nad doktoratem dr Anny Leśniewskiej. W tym czasie były dostępne pierwsze dane z sekwencjonowania całości RNA transkryptomu z próbek biologicznych. We współpracy z grupa prof. Schaffera ze Szpitala Dziecięcego w Zurichu analizowaliśmy za pomocą odczytów z sekwencera SOLiD ekspresjonowane RNA z linii komórkowych nowotworu rhybosarcoma (mięsak prążkowanokomórkowy). Ponieważ nie było wtedy wielu narzędzi programowych do analizy danych sekwencjonowania RNA, a większość z nich koncentrowała się na sprowadzaniu analizy do tabeli zliczeń odczytów w genach i analizy statystycznej podobnej jak w mikromacierzach [12, 13], postanowiliśmy na własne potrzeby stworzyć pakiet oprogramowania rna-

SeqMap **[S2]**, opisany w pracy **[P7]**. Pakiet rnaSeqMap zawierał obiekty pozwalające na przechowywanie i analizę danych z sekwencjonowania RNA dla określonych regionów genomowych dla wielu próbek biologicznych na raz. Dodatkowo możliwe było wykonywanie operacji takich jak zliczenia w regionach, wyznaczanie funkcji pokrycia genomu przez odczyty sekwencera, czy indeks splicingowy podobny do opisanego w [11], lecz obliczany dla każdego nukleotydu w regionie genomowym. Pakiet posiadał także elementy algorytmów uczenia maszyn, w szczególności implementację algorytmu Lindella-Aumanna. Algorytm ten odkrywający jednowymiarowe, nieredukowalne ilościowe reguły asocjacyjne stosowałem i modyfikowałem w swojej pracy doktorskiej [14], tym razem był on użyty do znajdowania bez nadzoru regionów o dużym pokryciu (eskpresji RNA) w funkcji pokrycia określonej na dziedzinie nukleotydów z regionu. W 2010 pakiet posiadał wiele funcjonalności komplementarnych do pierwszych wersji pakietu IRanges, czy też późniejszego Genomic Ranges.

Praca nad analizą funkcji pokrycia genomu odczytami z sekwencera dla RNA była kontynuowana w projekcie grantowym **[G2]**, którego wyniki stanowiły podstawę doktoratu dr Alicji Szabelskiej. Z wykorzystaniem interfejsu programistycznego biblioteki rnaSeqMap, zdefiniowane zostały nowe miary eskpresji różnicowej RNA, określone na regionach genomowych a obliczane z wykorzystaniem operacji na funkcjach pokrycia. Miary te, wraz z ich porównaniem i propozycjami zastosowań zostały opisane w pracy **[P8]**. W szczególności mogą służyć one do odkrywania nowych zjawisk splicingu alternatywnego w obrębie znanych regionów eksonów. Wielu biologów uważa eksony w bazach anotacyjnych za niepodzielne, zaś ten rodzaj analizy pokrycia pozwala na znalezienie eksonów, które mają nowe nieznane warianty, co może mieć wpływ białka otrzymane z nich poprzez translację i na ich znaczenie w procesach biologicznych.

Wyniki pracy **[P8]** zostały omówione na konferencji European Conference for Computational Biology w ramach całodziennego tutorialu na temat analiz sekwencjonowania RNA, jaki prowadziłem z prof. W. Huberem, prof. M. Robinsonem i dr S. Andersem (autorami prac [12, 13]). Podsumowana i uporządkowana wersja tego tutorialu ukazała się jako **[P10]** w czasopiśmie Nature Protocols, które definiuje standardy protokołów laboratoryjnych i analizy danych. Zawiera ona opis wraz z kodem w języku R do przetwarzania danych z sekwencjonowania RNA na poziomie podsumowań na genach. Przetwarzanie danych rozpoczyna się od plików w formacie FASTQ, zawierających odczyty z sekwencera wraz z miarami jakości dla każdego z nukleotydów w postaci współczynnika PHRED. Odczyty są mapowane do genomu za pomocą oprogramowania mapującego (*aligner*), które w wersji przeznaczonej dla RNA (np. pakiety tophat lub STAR) potrafią rozbić dopasowanie odczytu do więcej niż jednego eksonu i zapisać dopasowanie w postaci kodu CIGAR z przerwami na introny. Następnie odczyty dopasowane do regionów genowych są zliczane dla każdej próbki biologicznej - np.

metodą HTSeq S. Andersa [9]. W tabeli zliczeń kolumny odpowiadają próbkom biologicznym przypisanym do grup eksperymentalnych (*treatments*), a rzędy - genom. Aby porównać grupy eksperymentalne nie można zastosować testu t-Studenta, gdyż rozkłady zliczeń w genach nie mają rozkładu normalnego. Dlatego też metody DESeq [12] i edgeR [13] przyjmują, że jest to ujemny rozkład dwumianowy (Pascala) i aby wyznaczyć geny różnicujące stosują własne implementacje testów na modelu regresji ujemnej dwumianowej, różniące się przede wszystkim sposobami estymacji parametru dyspersji. W obecnych wersjach biblioteki DESeq i edgeR pozwalają na testowanie hipotez dla bardziej skomplikowanych układów eksperymentalnych niż porównywanie dwu grup próbek.

Uzupełnieniem algorytmów znajdowania ekspresji różnicowej za pomocą testów DESeq i edgeR są metody analizy ekspresji na poziomie eksonów takie jak DEXseq [15] czy metody analiz pokrycia opisane w **[P8]**.

Praca **[P10]** ma duże szanse stać się standardem opisującym sposób analizy danych z sekwencjonowania RNA. Jednakże wciąż jest miejsce na zwiększenie precyzji otrzymanych wyników i inteligentne znajdowanie regionów ekspresji, co pozwala na znajdowanie nowych wariantów splicingowych i nowych rodzajów RNA, np. długich niekodujących lincRNA, mogących mieć znaczenie regulatorowe w procesach komórkowych. Moja rola w przygotowaniu pracy **[P10]** dotyczyła na początku samego pomysłu standaryzacji przetwarzania danych z sekwencjonowania RNA i koordynowanie współpracy konkurencyjnych grup biostatystycznych prof. Hubera z Heidelbergu i prof. Robinsona z Zurichu. W dalszej części pomagałem w przygotowaniu danych testowych i uzupełnieniu metod o możliwości odkrywania alternatywnego splicingu i innych zjawisk nie pokrywających się z regionami genomowymi opisanymi w bazach anotacyjnych, wymagających metod uczenia bez nadzoru.

Należy dodać, że współpraca z grupami profesorów Hubera i Robinsona rozpoczęła się od działalności w ramach społeczności programistów repozytorium Bioconductor [10]. Jest to repozytorium bibliotek oprogramowania bioinformatycznego tworzonego w języku R na potrzeby nowych pod-dziedzin wiedzy łączących biologię z informatyką takich jak genomika, transkryptomika, proteomika czy metabolomika. Bioconductor jest administrowany przez grupę z Fred Hutchinson Cancer Research Center, która do roku 2009 prowadzona była przez Roberta Gentlemana, jednego z twórców języka R. Społeczność Bioconductora organizuje corocznie warsztaty i konferencje dla programistów - jeden w USA, drugi w Europie oraz wiele specjalizowanych kursów bioinformatycznych w różnych ośrodkach na całym świecie. W roku 2012 razem z profesorem Robinsonem byłem głównym organizatorem euro-

[9] http://www-huber.embl.de/users/anders/HTSeq
[10] http://www.bioconductor.org

pejskiego spotkania Bioconductora [11] W ramach organizacji tego spotkania otrzymaliśmy dofinansowanie na 20 grantów podróżnych z których 8 udało mi się przyznać dla polskich doktorantów i naukowców programujących w R.

W ramach grantów [**G3**] i [**G4**] kontynuuję badania nad precyzyjną analizą z dokładnością do nukleotydów dla sekwencjonowanego RNA. Pierwszą częścią było stworzenie pakietu oprogramowania ampliQueso [**S3**] do analizy danych z sekwencjonowania RNA ograniczonego do wielu krótkich amplikonów. Testy wykonane zostały na danych otrzymanych z bibliotek sekwencjonowania w technologii AmpliSeq (LifeTech) w której zakodowano ok 300 par primerów namnażających RNA z użyciem standardowej maszyny PCR. Pozwala to na zakodowanie za pomocą regionów amplikonów większego biomarkera - w naszym przypadku były to geny pozwalające rozróżniać pacjentów ze stwardnieniem rozsianym od zdrowych na podstawie próbek krwi. Wyniki w technologii AmpliSeq przypominają zbiór danych z małej mikromacierzy lub panelu PCR TaqMan. Odczyty są sekwencjonowane za pomocą sekwencerów IonTorrent lub IonProton. W odróżnieniu od QPCR czy mikromacierzy nie są analogowe lecz dyskretne, gdyż pochodzą ze zliczeń sekwencjonowanych odczytów. Dodatkową zaletą jest możliwość analizy polimorfizmów jednonukleotydowych (SNP) w amplikonach z dobrym pokryciem odczytami.

3.5 Precyzyjna analiza dużych mutacji DNA z użyciem sekwencerów wszystkich trzech generacji

Określanie różnic w DNA genomu takich jak mutacje w postaci insercji, delecji oraz polimorfizmy jednonukleotydowe również jest dziedziną w której można użyć połączenia dużej ilości danych generowanych z sekwencerów z ich analizą z nukleotydową precyzją. We współpracy z Gaborem Matyasem i Janine Meienberg z Centrum Genetyki Chorób Krążenia i Diagnostyki Genowej należącym do Fundacji na rzecz osób z rzadkimi chorobami w Zurichu-Schlieren zajęliśmy się tematyką precyzyjnego wyznaczenia zakresu rozległych (rzędu dziesiątek tysięcy par zasad) delecji odpowiedzialnych za choroby takie jak np. syndrom Marfana. Część delecji u pacjentów z takimi syndromami jest diagnozowana za pomocą sekwencjonowania Sangera. Szczególnie trudne do określenia ich położenia względem genomu referencyjnego są te, które na jednym lub obu swoich sekwencjach flankujących posiadają ciąg polinukleotydowy (np. poly-C, etc) lub region o małej złożoności - np. ciąg jednakowych par nukleotydów (np. GCGCGC...). Postanowiliśmy zatem sprawdzić jakie są różnice i możliwości identyfikacji delecji za pomocą sekwencjonowania Sangera, sekwencerów nowej generacji (Illumina HiSeq) i trzeciej generacji (PacBio). Każdy z trzech rodzajów sekwencerów

[11] http://www.fgcz.ch/Bioconductor2012/Overview

ma specyficzne rodzaje błędów i artefaktów na odczytach. Rozległe delecje są trudne do wyznaczenia metodą Sangera, wymagają dużego nakładu czasu i pracy oraz zastosowania wielu sekwencji primerów. Sekwencjonowanie DNA z użyciem Illumina HiSeq i PacBio pozwala na dokładne określenie zakresu delecji w ograniczonym czasie. Szczególnie efektywne jest zastosowanie PacBio, gdyż granice delecji są widoczne na funkcji pokrycia praktycznie natychmiast po mapowaniu do genomu referencyjnego. Jest ono również tańsze jeśli chodzi o ceny całego procesu sekwencjonowania, jednorazowych części i odczynników. Całość opisana została w pracy [**P9**]. Zawiera ona metodę algorytmiczno-statystyczną znajdowania precyzyjnej ilości motywów polinukleotydowych lub powtarzalnych w sekwencjach flankujących. Została ona opracowana przeze mnie wraz z dr Alicją Szabelską w ramach jej pracy w projekcie [**G2**].

3.6 Precyzyjne i skalowalne analizy danych genomowych w środowiskach chmur obliczeniowych

Obecne technologie odczytywania DNA i RNA pozwalają na precyzyjne odczytywanie ich fragmentów, z dokładnością do pojedynczego nukleotydu. Kolejne generacje tych technologii, od mikromacierzy 3' poprzez mikromacierze eksonowe do sekwencjonowania nowej generacji dążyły do jak największej precyzji pomiaru i uzyskania jak największej ilości danych pozwalających na odczytanie jednocześnie wielu regionów genomowych czy transkryptów. W chwili obecnej odczytanie całości genomu czy transkryptomu jest możliwe, barierą są co najwyżej koszty eksperymentu. Generowane są przy tym duże ilości danych, wymagające nowych metod i algorytmów przetwarzania, ale też pozwalające na nowe zastosowania w biologii i medycynie. Przykładami najnowszych problemów rozwiązywalnych za pomocą tych technik może być np. badanie metylacji genomu, znajdowanie nowych form i sekwencji regulatorowego RNA czy odkrywanie mechanizmów regulacji procesów komórkowych poprzez alternatywny splicing genów. Dalsze zwiększenie jakości otrzymywanych wyników w genomice i transkryptomice będzie możliwe, jednakże wymagać będzie ono analiz coraz większych ilości danych z sekwencjonowania, co z kolei wymaga stosowania dużych pojemności dysków i metod algorytmów równoległych. W tym kierunku idą moje obecne badania, objęte grantami [**G3**] , [**G4**] i [**G5**] dotyczące zastosowania chmur obliczeniowych i systemów operowania w nich z użyciem paradygmatów mapowania-redukcji takich jak Apache Spark i narzędzi z "ekosystemu Hadoop" do znajdowania zjawisk w genomach i transkryptomach z dokładnością do pojedynczych nukleotydów. Prototypowe oprogramowanie SparkSeq opisane w [**P11**] jest wynikiem tych badań, udowadnia ono możliwość skalowalności obliczeń genomowych z użyciem Apache Spark i jest o rząd wielkości szybsze obliczeniowo od niedawnych rozwiązań takich jak np. SeqPig [16].

Potwierdzeniem właściwego wyboru kierunku badań jest na przykład fakt ukazania się niedawno artykułu [17] będącego opisem nowej metody znajdowania regionów ekspresji różnicowej z nukleotydową precyzją, inspirowany częściowo pracą **[P7]**, a powstały w grupach dwu ważnych autorytetów biostatystyki, prof. Irizarry'ego i prof Leek'a.

3.7 Podsumowanie

Artykuły **[P1]**- **[P11]** stanowią opis konsekwentnego dążenia do otrzymania jak najbardziej użytecznych dla biologii i medycyny wyników analiz genomowych. Praca nad nimi odbywała się w niezwykle ciekawym okresie rozwoju nanotechnologii i technik obliczeniowych, powodujących stałe udoskonalanie narzędzi sprzętowych w dziedzinie transkryptomiki i genomiki.

Pozostałe moje publikacje, nie wchodzące bezpośrednio w skład osiągnięcia naukowego opisanego w tytule, stanowią w większości zastosowania metod bioinformatycznych wchodzących w skład rozwiązania określonych problemów biologii molekularnej czy medycyny. Często te metody bioinformatyczne zawierają elementy opracowane przeze mnie w głównym toku mojej pracy naukowej. Chodzi tu np. o artykuły wykorzystujące analizę mikromacierzy eksonowych do znajdowania zbiorów genów różnicowo ekspresjonowanych i poznania procesów zachodzących w komórkach i tkankach nowotworowych, wyjaśnienia procesów kierowanych "nonsensownymi" transkryptami genów czy porównania procesów wspólnych dla embrionalnych komórek macierzystych w różnych gatunkach.

Od strony zastosowania paradygmatów informatyki, główną z myśli przewodnich mojej pracy naukowej było wykorzystanie danych jak najbardziej surowych a jak najmniej zagregowanych co zapewnia w procesie odkrywania wiedzy największą precyzję wyników, choć może wymagać większych mocy obliczeniowych i bardziej specyficznych rozwiązań oprogramowania. Takiego podejścia do analizy danych nauczyłem się podczas studiów doktoranckich na Politechnice Warszawskiej, gdzie zajmowałem się algorytmami odkrywania wiedzy pod kierunkiem prof. Muraszkiewicza i prof. Rybińskiego. Ten sposób myślenia ukierunkował moje badania po doktoracie i zaowocował pracami składającymi się na osiągnięcie naukowe opisane w niniejszym autoreferacie. Również wiele z metod stworzonych na potrzeby artykułów wchodzących w skład osiągnięcia naukowego powstało na bazie rozwijania i łączenia metod z dziedzin takich jak bazy danych, uczenie maszyn, automatyczne odkrywanie wiedzy i statystyka. Wszystkie te z nich rozpocząłem poznawać podczas moich studiów magisterskich i doktoranckich na Politechnice Warszawskiej i wiedza ta okazała się niezwykle przydatną podstawą w zastosowaniach bioinformatycznych w biologii molekularnej i medycynie.

4 Ocena wkładu pracy w publikacje należące do osiągnięcia naukowego

Pozycja nazwisk autorów odzwierciedla system stosowany w naukach biologicznych i medycynie. Pierwszy autor jest osobą odpowiedzialną za wykonanie największego wkładu pracy badawczej, ostatni "senior" autor jest osobą, która otrzymała finansowanie na pracę badawczą i nadzorującą ją. Przy równym wkładzie pracy pierwszych autorów, są oni oznaczani jako "wspólny pierwszy autor".

[P1] pierwszy autor, wkład 60%

Opis wkładu pracy umieszczony w artykule: MO developed the concept of interactions in families of probesets, carried out database and statistical analyses and drafted the manuscript, CM conceived the study on probes alignments to transcript and its implications, supervised and participated in its design and helped to draft the manuscript.

[P2] pierwszy autor, wkład 60%

[P3] pierwszy autor, wkład 50%

[P4] współautor, wkład 30%

[P5] pierwszy autor, wkład 45%

Opis wkładu pracy umieszczony w artykule: CJM conceived and designed and performed the experiments, analyzed the data, contributed reagents/materials/analysis/tools, and wrote the paper. MJO performed the experiments, analyzed the data, and contributed reagents/materials/analysis/tools.

[P6] pierwsze autorstwo dzielone, wkład 30%

Opis wkładu pracy umieszczony w artykule: DB and MO performed the data analysis. YC performed the proteomics bench work. CM conceived the experiment and wrote the manuscript.

[P7] "senior" autor, wkład 40%

Opis wkładu pracy umieszczony w artykule: AL has prepared the implementation of rnaSeqMap, developed the algorithms, heuristics and the adaptation of xmapcore database, performed the quantitative experiments and helped to draft the manuscript. MO conceived the idea of the software, participated in the implementation of rnaSeqMap and drafted the manuscript. Both authors read and approved the final manuscript.

[P8] pierwszy autor, wkład 40%

[P9] pierwsze autorstwo dzielone, wkład 40%

[P10] współautor, wkład 10%

Opis wkładu pracy umieszczony w artykule: S.A. and W.H. are authors of the DESeq package. D.J.M., Y.C., G.K.S. and M.D.R. are authors of the edgeR package. S.A., M.O., W.H. and M.D.R. initiated the protocol format on the basis of the ECCB 2012 Workshop. S.A. and M.D.R. wrote the first draft and additions were made from all authors

[P11] "senior" autor, wkład 40%

5 Spis publikacji naukowych po doktoracie nie wchodzących w skład osiągnięcia naukowego

[D1] Surace, L., Lysenko, V., Fontana, A. O., Cecconi, V., Janssen, H., Bicvic, A., **Okoniewski MJ**, Pruschy M., Dummer R., Neefjes J., Knuth A., Gupta A., van den Broek, M.. Complement is a central mediator of radiotherapy-induced tumor-specific immunity and clinical response. Immunity, 2015, 42(4), 767-777.

IF 19.75,

[D2] Meienberg J., Zerjavic K., Keller I., **Okoniewski MJ**, Patrignani A., Ludin K., et al. . New insights into the performance of human whole-exome capture platforms. Nucleic Acids Research, 2015, gkv216.

IF 8.81,

[D3] Walker RA., Sharman PA., Miller CM., Lippuner C., **Okoniewski MJ**, Eichenberger RM., et al. RNA Seq analysis of the Eimeria tenella gametocyte transcriptome reveals clues about the molecular basis for sexual reproduction and oocyst biogenesis. BMC Genomics, 2015, 16(1), 94. doi:10.1186/s12864-015-1298-6

IF 4.04,

[D4] Hehl AB., Basso, WU., Lippuner C., Ramakrishnan C., **Okoniewski MJ**, Walker RA., et al. Asexual expansion of Toxoplasma gondii merozoites is distinct from tachyzoites and entails expression of non-overlapping gene families to attach, invade, and replicate within feline enterocytes. BMC Genomics, 2015, 16(1), 66. doi:10.1186/s12864-015-1225-x

IF 4.04,

[D5] Urosevic-Maiwald M., Barysch MJ., Cheng PF., Karpova MB., Steinert H., **Okoniewski MJ**., Dummer R. . In vivo profiling reveals immuno-

modulatory effects of sorafenib and dacarbazine on melanoma. OncoImmunology, 2014
IF 6.28,

[D6] Draganova, K., Zemke, M., Zurkirchen L., Valenta T., Cantu C., **Okoniewski MJ**., et al.. Wnt/beta-catenin signaling regulates sequential fate decisions of murine cortical precursor cells. Stem Cells 2015, 33(1), 170–182. doi:10.1002/stem.1820
IF 7.07,

[D7] Allie F., Pierce EJ., **Okoniewski MJ**., Rey C. Transcriptional analysis of South African cassava mosaic virus-infected susceptible and tolerant landraces of cassava highlights differences in resistance, basal defense and cell wall associated genes during infection. BMC Genomics, 2014 15, 1006. doi:10.1186/1471-2164-15-1006
IF 4.04,

[D8] Wachtel M, Rakic J, **Okoniewski MJ**, Bode P, Niggli F, Schäfer BW, FGFR4 signaling couples to bim and not bmf to discriminate subsets of alveolar rhabdomyosarcoma cells. International Journal of Cancer, 2014 Feb 18. doi: 10.1002/ijc.28800
IF 6.19,

[D9] Vonlanthen J, **Okoniewski MJ**, Meningatti M, Cattaneo E, Pellegrini-Ochsner D, Haider R, Jiricny J, Staiano T, Buffoli F, Marra G., A comprehensive look at transcription factor gene expression changes in colorectal adenomas. BMC Cancer 2014, 14:46 doi:10.1186/1471-2407-14-46
IF 3.33,

[D10] Sharma A, Bode B, Moch H, **Okoniewski M**, Knuth A, von Boehmer L, van den Broek M, Radiotherapy of human sarcoma promotes an intratumoral immune effector signature, Clinical Cancer Research, 2013 Epub Jul 16.
IF 7.84,

[D11] Tiwari A, Schneider M, Fiorino A, Haider R, **Okoniewski MJ**, Roschitzki B, Uzozie A, Menigatti M, Jiricny J, Marra G., Early Insights into the Function of KIAA1199, a Markedly Overexpressed Protein in Human Colorectal Tumors, PLoS One. 2013 Jul 23;8(7):e69473. doi: 10.1371/journal.pone.0069473.
IF 3.73,

[D12] Mohme M, Hotz C, Stevanovic S, Binder T, Lee JH, **Okoniewski M**, Eiermann T, Sospedra M, Rammensee HG, Martin R. HLA-DR15-

derived self-peptides are involved in increased autologous T cell proliferation in multiple sclerosis., Brain, 2013 Jun;136(Pt 6):1783-98. doi: 10.1093/brain/awt108

IF 9.92,

[D13] Casanova EA*, **Okoniewski MJ***, Cinelli P, Cross-Species Genome Wide Expression Analysis during Pluripotent Cell Determination in Mouse and Rat Preimplantation Embryos. PLoS One. 2012;7(10):e47107.

IF 3.73,

[D14] Shakhova O, Zingg D, Schaefer SM, Hari L, Civenni G, Blunschi J, Claudinot S, **Okoniewski M**, Beerman F, Mihic-Probst D, Moch H, Dummer R, Barrandon Y, Cinelli P, Sommer L. Sox10 promotes the formation and maintenance of giant congenital naevi and melanoma. Nature Cell Biology, 14(8), 2012

IF 20.76,

[D15] Egger A, Samardzija M, Sothilingam V, Tanimoto N, Lange C, Salatino S, Fang L, Garcia-Garrido M, Beck S, **Okoniewski MJ**, Neutzner A, Seeliger MW, Grimm C, Handschin C, PGC-1alpha Determines Light Damage Susceptibility of the Murine Retina, PLoS One. 2012;7(2):e31272. Epub 2012 Feb 13

IF 3.73,

[D16] Yepiskoposyan H, Aeschimann F, Nilsson D, **Okoniewski M**, Mühlemann O, Autoregulation of the nonsensemediated mRNA decay pathway in human cells, RNA Journal. 2011

IF 5.09,

[D17] Kariminejad A, Kariminejad R, Moshtagh A, Zanganeh M, Kariminejad MH, Neuenschwander S, **Okoniewski MJ**, Wey E, Schinzel A, Baumer A., Pericentric inversion of chromosome 18 in parents leading to a phenotypically normal child with segmental uniparental disomy 18, Eur J Hum Genet. 2011 May;19(5):555-60.

IF 3.56,

[D18] Imig J, Motsch N, Zhu JN, Barth S, **Okoniewski MJ**, Reineke T, Tinguely M, Faggioni A, Trivedi P, Meister G, Renner Ch, Grasser F, MicroRNA profiling in EBV-associated B-cell lymphoma, Nucleic Acids Research, 2010

IF 8.81,

[D19] Sims AH, Smethurst GJ, Hey Y,**Okoniewski MJ**, Pepper SD, Howell A, Miller CJ, Clarke RB. The removal of multiplicative, systematic

bias allows integration of breast cancer gene expression datasets – improving metaanalysis and prediction of prognosis. BMC Medical Genomics 2008, 1:42
IF 3.47,

[D20] Pierce A, Unwin RD, Evans CA, Griffiths S, Carney L, Zhang L, Jaworska E, Lee CF, Blinco D, **Okoniewski MJ**, Miller CJ, Bitton DA, Spooncer E, Whetton AD. Eight-channel iTRAQ enables comparison of the activity of six leukemogenic tyrosine kinases. Mol Cell Proteomics. 2008 May;7(5):853-63.
IF 7.25,

Literatura dodatkowa

[1] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977.

[2] E S Lander, L M Linton, B Birren, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[3] T W Chang. Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of immunological methods*, 65(1-2):217–223, December 1983.

[4] M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235):467–470, October 1995.

[5] Philipp Kapranov, Simon E Cawley, Jorg Drenkow, Stefan Bekiranov, Robert L Strausberg, Stephen P A Fodor, and Thomas R Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science (New York, N.Y.)*, 296(5569):916–919, May 2002.

[6] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4):e15, February 2003.

[7] Hui Sun Leong, Tim Yates, Claire Wilson, and Crispin J Miller. ADAPT: a database of affymetrix probesets and transcripts. *Bioinformatics (Oxford, England)*, 21(10):2552–2553, May 2005.

[8] Manhong Dai, Pinglang Wang, Andrew D Boyd, Georgi Kostov, Brian Athey, Edward G Jones, William E Bunney, Richard M Myers, Terry P

Speed, Huda Akil, et al. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic acids research*, 33(20):e175–e175, 2005.

[9] Lev Klebanov and Andrei Yakovlev. Diverse correlation structures in gene expression data and their utility in improving statistical inference. *The Annals of Applied Statistics*, 1(2):538–559, 12 2007.

[10] Roman Jaksik, Joanna Polańska, Robert Herok, and Joanna Rzeszowska-Wolny. Calculation of reliable transcript levels of annotated genes on the basis of multiple probe-sets in affymetrix microarrays. *Acta Biochimica Polonica*, 56(2):271, 2009.

[11] Paul Gardina, Tyson Clark, Brian Shimada, Michelle Staples, Qing Yang, James Veitch, Anthony Schweitzer, Tarif Awad, Charles Sugnet, Suzanne Dee, Christopher Davies, Alan Williams, and Yaron Turpaz. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, 7(1):325, 2006.

[12] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

[13] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, January 2010.

[14] Michał J Okoniewski. *Odkrywanie wielowymiarowych ilościowych reguł asocjacyjnych*. PhD thesis, Politechnika Warszawska, 2002.

[15] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, October 2012.

[16] André Schumacher, Luca Pireddu, Matti Niemenmaa, Aleksi Kallio, Eija Korpelainen, Gianluigi Zanetti, and Keijo Heljanko. Seqpig: simple and scalable scripting for large sequencing data sets in hadoop. *Bioinformatics*, 30(1):119–120, 2014.

[17] Alyssa C Frazee, Sarven Sabunciyan, Kasper D Hansen, Rafael A Irizarry, and Jeffrey T Leek. Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics (Oxford, England)*, January 2014.

Uzupełnienie dokumentu Autoreferatu
3 sierpnia 2015

Ocena wkładu pracy w publikacje nalezace do osiagniecia naukowego

Pozycja nazwisk autorów odzwierciedla system stosowany w naukach biologicznych i medycynie. Pierwszy autor jest osoba odpowiedzialna za wykonanie najwiekszego wkładu pracy badawczej, ostatni "senior" autor jest osoba, która otrzymała finansowanie na prace badawcza i nadzorujaca ja. Przy równym wkładzie pracy pierwszych autorów, sa oni oznaczani jako "wspólny pierwszy autor".

[P1] pierwszy autor, wkład 60%
Opis wkładu pracy umieszczony w artykule: MO developed the concept of interactions in families of probesets, carried out database and statistical analyses and drafted the manuscript, CM conceived the study on probes alignments to transcript and its implications, supervised and participated in its design and helped to draft the manuscript.

Wkład MO polegał na stworzeniu koncepcji badań opisanej w artykule, analizach statystycznych i baz danych oraz wspólnym napisaniu manuskryptu.

[P2] pierwszy autor, wkład 60%

Wkład MO polegał na stworzeniu koncepcji badań opisanej w artykule, analizach statystycznych i baz danych oraz wspólnym napisaniu manuskryptu.

[P3] pierwszy autor, wkład 50%

Wkład MO polegał na stworzeniu koncepcji badań opisanej w artykule, analizach danych oraz wspólnym napisaniu manuskryptu

[P4] współautor, wkład 30%

Wkład MO polegał na współtworzeniu koncepcji badań opisanej w artykule, analizach statystycznych i baz danych oraz wspólnym napisaniu manuskryptu.

[P5] pierwszy autor, wkład 45%
Opis wkładu pracy umieszczony w artykule: CJM conceived and designed and performed the experiments, analyzed the data, contributed reagents/ materials/analysis/tools, and wrote the paper. MJO performed the experiments, analyzed the data, and contributed reagents/materials/analysis/tools.

Wkład MO polegał na przeprowadzeniu eksperymentów, analizie danych i wkładzie w analize i narzędzia badawcze.

[P6] pierwsze autorstwo dzielone, wkład 30%
Opis wkładu pracy umieszczony w artykule: DB and MO performed the data analysis. YC performed the proteomics bench work. CM conceived the experiment and wrote the manuscript.

Wkład MO polegał na przeprowadzeniu analizy danych.

[P7] "senior" autor, wkład 40%
Opis wkładu pracy umieszczony w artykule: AL has prepared the implementation of rnaSeqMap, developed the algorithms, heuristics and the adaptation of xmapcore database, performed the quantitative experiments and helped to draft the manuscript. MO conceived the idea of the software, participated in the implementation of rnaSeqMap and drafted the manuscript. Both authors read and approved the final manuscript.

Wkład MO polegał na stworzeniu idei oprograowania, udziale w jego implementacji i napisaniu manuskryptu pracy.

[P8] pierwszy autor, wkład 40%

Wkład MO polegał na stworzeniu idei metody i oprograowania, udziale w jego implementacji i napisaniu manuskryptu pracy.

[P9] pierwsze autorstwo dzielone, wkład 40%

Wkład MO polegał na stworzeniu koncepcji badań, analizie danych i udziale w napisaniu manuskryptu pracy.

[P10] współautor, wkład 10%
Opis wkładu pracy umieszczony w artykule: S.A. and W.H. are authors of the DESeq package. D.J.M., Y.C., G.K.S. and M.D.R. are authors of the edgeR package. S.A., M.O., W.H. and M.D.R. initiated the protocol format on the basis of the ECCB 2012 Workshop. S.A. and M.D.R. wrote the first draft and additions were made from all authors

Wkład MO polegał na rozpoczeciu tworzenia protokołu analizy na baize materiałów stworzonych ze współautorami na workshop konferencji ECCB.

[P11] "senior" autor, wkład 40%

Wkład MO polegał na stworzeniu koncepcji i planu oprograowania, udziale w jego testowym zastosowaniu i napisaniu manuskryptu pracy.

Dodatek 2 do Autoreferatu
Aug 24th 2015

Ocena wkładu w pozostałe prace naukowe

[D1] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 5%
[D2] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 20%
[D3] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 25%
[D4] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 25%
[D5] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 10%
[D6] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 10%
[D7] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 25%
[D8] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 10%
[D9] analiza danych genomowych, przygotowanie manuskryptu, współautor koncepcji pracy
ocena wkładu: 40%
[D10] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 5%
[D11] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 15%
[D12] statystyczna analiza danych klinicznych, pomoc w przygotowaniu manuskryptu ocena wkładu: 10%
[D13] analiza danych genomowych, przygotowanie manuskryptu, współautor koncepcji pracy
ocena wkładu: 40%
[D14] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 5%
[D15] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 10%
[D16] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 25%
[D17] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 5%
[D18] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 5%
[D19] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 10%
[D20] analiza danych genomowych, pomoc w przygotowaniu manuskryptu
ocena wkładu: 10%

25.08.15