ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ
Faculty of Biology
Institute of Human Biology and Evolution

prof. dr hab. Izabela Makałowska
Instytut Biologii I Ewolucji Człowieka
Wydział Biologii
Uniwersytet im. Adama Mickiewicza w Poznaniu

Poznań, 3.04.2023

Recenzja pracy doktorskiej mgr Bence Galika, doktoranta Uniwersytetu Medycznego w Białymstoku, pt „Opracowanie i wdrożenie oryginalnych bioinformatycznych protokołów analitycznych do analizy danych NGS w badaniach nad nowotworami i rozrodem wspomaganym."

W ostatnich latach obserwujemy prawdziwą rewolucję w badaniach prowadzonych w obszarze nauk przyrodniczych, w tym oczywiście także badań biomedycznych. Wynika to głównie z rozwoju nowych, wysokoprzepustowych technologii sekwencjonowania, które uczyniły genomikę i transkryptomikę, wraz z innymi badaniami -omicznymi, integralną częścią badań medycznych. Masowość i rosnąca złożoność projektów, w których wykorzystuje się sekwencjonowanie nowej generacji stawia przed bioinformatykami nowe wyzwania. Analiza wyników szeroko zakrojonych badań -omicznych często wymusza jednoczesne stosowanie wielu, niekoniecznie kompatybilnych, narzędzi bioinformatycznych. Różnorodność typów danych oraz specyfika poszczególnych badań utrudniają lub wręcz uniemożliwiają zaprojektowanie uniwersalnych protokołów do przetwarzania i analizy danych. Niemal każdy projekt wymaga specyficznej dla projektu ścieżki analitycznej. Zaprojektowanie i implementacja ścieżek do analizy danych z wysokoprzepustowego sekwencjonowania oraz wykonanie wymaganych analiz było głównym celem rozprawy przygotowanej przez Bence Galika. Przedstawione badania koncentrowały się na pięciu przypadkach dotyczących przewlekłej białaczki limfatycznej, pierwotnego chłoniaka ośrodkowego układu nerwowego, glejaka wielopostaciowego, niedrobnokomórkowego raka płuca oraz preimplantacyjnych badań genetycznych zarodków w zapłodnieniu in vitro. Praca została wykonana pod kierunkiem dr hab. Attila Gyenesei w Zakładzie Klinicznej Biologii Molekularnej Uniwersytetu Medycznego w Białymstoku.

Ocena formalna rozprawy

Ocena przedstawionej rozprawy pod względem formalnym była dla mnie nico kłopotliwa. Wynika to przede wszystkim z tego, że przeprowadzona analiza antyplagiatowa wykazała 61% podobieństwa do istniejących dokumentów. Dalsza analiza jednoznacznie wykazała, że cztery przypadki badawcze spośród pięciu składających się na rozprawę były opublikowane. Części opisujące te badania to przepisane, z zasadniczo nieistotnymi zmianami, akapity zaczerpnięte z tych publikacji. Stanowi to prawie 80% całej rozprawy. Samo wykorzystanie opublikowanych

ul. Uniwersytetu 61-614 Poznań, Poland
tel. +48 61 829 57 30, + 48 61 829 58 35
anthro@amu.edu.pl
www.biologia.amu.edu.pl

wyników nie stanowi problemu. Skala jest jednak dość duża, a Doktorant nie wspomniał ani o tym, że wyniki zostały opublikowane, ani też o innych osobach biorących udział w badaniach. W ten sposób niejako przypisał sobie całą wykonaną pracę. Budzi to wątpliwości, gdyż wszystkie publikacje były pracami wieloautorskimi. Dlatego też przed przystąpieniem do oceny poprosiłam doktoranta o szczegółowy opis wkładu w każde z opisywanych w publikacjach badań.

W trzech z tych publikacji mgr Galik jest równorzędnym pierwszym autorem i moim zdaniem przedstawienie rozprawy w formie zbioru tych trzech prac spełniałoby zarówno wymogi ustawowe, jak i te określone w uchwale Senatu Uniwersytetu Medycznego w Białymstoku z dnia 24 października 2019 r. (uchwała 91/2019, załącznik 1f). Pozwoliłoby to na uniknięcie wskazanych problemów. Chciałabym poprosić Doktoranta o wyjaśnienie tego nieoczywistego dla mnie wyboru sposobu przygotowania rozprawy.

Przedstawiona do recenzji rozprawa jest opracowaniem obejmującym 128 stron tekstu w języku angielskim, ilustrowanego 31 rysunkami i zawierającego 19 tabel. Spis cytowanej literatury jest obszerny i zawiera ponad 200 pozycji. Układ pracy jest typowy dla dysertacji doktorskich. Praca składa się ze streszczenia (w języku polskim i angielskim), wstępu, materiałów i metod, wyników oraz dyskusji. Całość zamykają zwięzłe konkluzje. Zabrakło natomiast jasno sprecyzowanego przez Doktoranta celu pracy. Praca została napisana dość poprawnie choć doktorant nie ustrzegł się pewnych niedoskonałości językowych i edytorskich. Podam jedynie kilka przykładów:

- na stronie 16 znajduje się odniesienie do [IJC], prawdopodobnie chodziło o odniesienie do którejś z pozycji literaturowych;

- na stronie 17 jest "Embry" zamiast "Embryo"

- na stronie 20 zamiast „in during the run..." powinno być "during the run ..."

- na stronie 57 jest napisane „Also, other interesting patient specific events were obese during the study". Prawdopodobnie miało być „...patient specific events were observed ..." a nie „were obese".

Styl pisania rozprawy nie jest spójny. Zdania są pisane czasem w czasie przeszłym, czasem w teraźniejszym, a czasem w przyszłym. Edytorskim niedopatrzeniem są także nieczytelne opisy na niektórych rysunkach, przykładowo na rysunku 12. Dodatkowo, wiele skrótów nie zostało wyjaśnionych, a praca nie zawiera spisu skrótów. W niektórych z opisanych badań brakuje informacji o pochodzeniu próbek.

Ocena merytoryczna

We wstępie Bence Galik skupił się przede wszystkim na omówieniu poszczególnych schorzeń, z którymi powiązane były analizy. Niestety niewiele jest o bioinformatyce, wysokoprzepustowych technologiach sekwencjonowania i ich zastosowaniach, a także o istniejących narzędziach. Zagadnienia te zostały omówione bardzo skrótowo i w mojej ocenie w sposób nie do końca zadowalający, szczególnie biorąc pod uwagę bioinformatyczny charakter rozprawy.

Pierwszy z badanych przypadków dotyczy przewlekłej białaczki limfatycznej (z ang. CLL). Celem przeprowadzonych badań było przeanalizowanie ewolucji klonalnej 30 genów, których mutacje są regularnie obserwowane w przypadkach CLL. Badanie przeprowadzono w kontekście terapii ibrutynibem w oparciu o celowane sekwencjonowanie genów będących przedmiotem zainteresowania. Próbki pochodziły z krwi obwodowej 20 pacjentów i pobrane zostały przed i po leczeniu. W wyniku przeprowadzonych analiz doktorant zidentyfikował łącznie 211 wariantów somatycznych i wykazał, że całkowita liczba mutacji w próbkach po

leczeniu wzrosła. Stwierdził również wielokrotne mutacje aż w 40% analizowanych genów. Co ciekawe, klony niosące mutacje w genach *IGLL5, EIF2A* i *EP300* zostały wyeliminowane w wyniku leczenia. Ponadto doktorant opisał nowy wzór dynamiki klonalnej mutacji w genach *BTK* i *TP53*. Uzyskane wyniki są bardzo interesujące. Pokazują, że kompleksowe, głębokie sekwencjonowanie genów i analiza ewolucji klonalnej mogą przynieść korzyści kliniczne. Jest to istotne dla lepszej charakterystyki choroby u poszczególnych pacjentów, a w konsekwencji lepszego, bardziej spersonalizowanego leczenia.

Drugi przypadek badawczy dotyczył pierwotnego chłoniaka ośrodkowego układu nerwowego (z ang. PCNSL). Podobnie jak w przypadku CLL wykonano celowane sekwencjonowanie amplikonów, z tą różnicą, że DNA ekstrahowano z próbek FFPE. Materiał pochodził od 64 pacjentów z pierwotnym chłoniakiem i 12 pacjentów z chłoniakiem wtórnym (SCNSL). Profile mutacji określone zostały przez doktoranta dla 14 genów. Przeprowadzona analiza wykazała 239 mutacji we wszystkich chłoniakach razem. 210 z nich zaobserwowano u pacjentów z pierwotnym chłoniakiem. Genami z największą liczbą mutacji w przypadku PCSNL były *MYD88, PIM1, KMT2D* i *PRDM1*. Nie zaobserwowano natomiast mutacji w genie *PTPRD*. W kohorcie pacjentów z wtórnym chłoniakiem mutacje najczęściej występowały w genach *PRDM1, MYD88 i PIM1*. W genach *CARD11, CSMD2, CSMD3* i *PTPRD* nie zidentyfikowano ani jednaj mutacji. Doktorant następnie porównał różnice w dystrybucji mutacji u pacjentów z pierwotnym i wtórnym chłoniakiem, a także pomiędzy podtypami ABC i GC. Nie znaleziono jednak różnic, które osiągnęłyby istotność statystyczną.

Mam dwa pytania do Doktoranta dotyczące tych badań:

1. W opisie analizy bioinformatycznej jest odniesienie do bibliotek specyficznych dla próbek A i B, które to biblioteki zostały połączone (str. 35). Brakuje jednak wyjaśnienia, czym są biblioteki A i B. Chciałbym, o to wyjaśnienie poprosić?

2. W opisie wyników znajduje się porównanie dwóch metod klasyfikacji pacjentów. W metodach opisana jest tylko jedna z nich – test LST. Algorytm Hansa został po raz pierwszy wymieniony w opisie wyników i nie został dobrze wyjaśniony. Rysunek przedstawiający ten algorytm jest opisany w sposób zbyt ogólny i niewystarczający dla osób nie będących ekspertami w klasyfikacji chłoniaków. Prosiłbym o wyjaśnienie tego algorytmu.

W kolejnych badaniach Bence Galik analizował dane dotyczące glejaka wielopostaciowego (GBM), który jest bardzo agresywnym guzem mózgu. Celem była identyfikacja molekularnych czynników powiązanych z rozwojem i nawrotami glejaka wielopostaciowego. Tym razem analiza dotyczyła wzorców metylacji DNA. Wykorzystana została technika sekwencjonowania bibliotek przekształconych wodorosiarczynem. Doktorant porównał wzorce metylacji w próbach kontrolnych, guzach przed radiochemioterapią i guzach nawracających po leczeniu. DNA pochodziło z próbek FFPE 22 pacjentów. Jako kontrolę wykorzystano dane zdeponowane w bazie EGA (The European Genome-phenome Archive). Były to próbki pobrane podczas operacji pacjentów z epilepsją. Porównanie poziomu metylacji wykazało przesunięcie w kierunku hipometylacji – największy poziom metylacji był w kontroli, najmniejszy w przypadku guzów nawracających po leczeniu. Niemniej, analiza różnicowa poziomu metylacji w odniesieniu do poszczególnych rejonów nie ujawniła istotnych statystycznie różnic. Natomiast analizy GO pozwoliły na identyfikację kilku istotnych ścieżek metabolicznych. W oparciu o uzyskane dane zaproponowany został mechanizm mogący leżeć u podstaw rozwoju i progresji glejaka wielopostaciowego. Uzyskane wyniki, a szczególnie zaproponowany mechanizm mogą być istotne dla opracowania nowych metod leczenia.

W powyższych analizach wykorzystano próbki z dwóch różnych źródeł co wprowadza zróżnicowanie pod względem technicznym, które może i najprawdopodobniej wpłynęło na

wyniki analizy. Czy oceniono wpływ tych różnic i zastosowano jakieś techniki pozwalające na zniwelowanie tego wpływu?

Czwarty projekt dotyczył ekspresji miRNA w niedrobnokomórkowym raku płuca (z ang. NSCLC). Analizie poddano próbki guza, krwi obwodowej i kontroli pochodzące od pacjentów ze zdiagnozowanym rakiem gruczołowym (z ang. AC) i rakiem płaskonabłonkowym płuca (z ang. SCC). Doktorantowi udało się zidentyfikować ponad 1600 miRNA. W celu porównania poszczególnych próbek i identyfikacji tych odstających obliczona została korelacja pomiędzy nimi. Podobieństwo próbek zostało również sprawdzone na podstawie hierarchicznego grupowania i analizy składowych głównych. Na dendrogramie próbki kontrolne wyraźnie oddzieliły się od próbek pochodzących z guza. Analiza składowych głównych wykazała zróżnicowanie pomiędzy próbkami gruczolakoraka i kontrolą jak i pomiędzy próbkami raka płaskonabłonkowego i jego kontroli. Co jednak ciekawe, próbki gruczolakoraka pokrywały się z kontrolą raka płaskonabłonkowego, a próbki raka płaskonabłonkowego z kontrolą gruczolakoraka. Czy można to w jakiś sposób wyjaśnić?

Dalsza analiza, różnicowa analiza ekspresji, pozwoliła doktorantowi na wyłonienie miRNA o zwiększonej i zmniejszonej w obu typach raka ekspresji, jak i miRNA o zmianach specyficznych dla danego typu. Z kolei analiza KEEG ujawniła szlaki metaboliczne powiązane z mikroRNA leżącymi u podstaw patogenezy raka gruczołowego i raka płaskonabłonkowego płuca. Ostatnim krokiem było zbudowanie modelu na podstawie 17 miRNA, które wykazywały największe różnice w ekspresji i potencjalnie mogą być biomarkerami. Model poddano walidacji na zbiorze danych pochodzących z próbek krwi. Wyniki walidacji nie były do końca satysfakcjonujące co zapewne wynika z chyba nie najlepiej dobranego zestawu danych. Ekspresja miRNA w krwi, z pewnością nie odpowiada tej w tkankach nowotworowych co zauważył także doktorant. Dlatego chciałam zapytać, dlaczego właśnie taki zestaw danych został wybrany w celu walidacji modelu?

Celem ostatniego projektu było opracowanie bioinformatycznego protokołu analitycznego na potrzeby nieinwazyjnych preimplantacyjnych badań genetycznych (NIPGT-A). Łącznie zsekwencjonowano 28 próbek, a 22 wybrano do dalszej analizy w oparciu o jakość sekwencjonowania. Analiza obejmowała pięć próbek pożywki kontrolnej, dwie próbki krwi pępowinowej ze znanymi wariantami liczby kopii (z ang. CNV) i 15 próbek kropli pożywki hodowlanej 3 dniowych zarodków. Sześć próbek pochodziło z medium zarodków poronnych, a dziewięć zdrowo urodzonych. Po sekwencjonowaniu genomowego DNA zidentyfikowano warianty liczby kopii. Analiza przeprowadzona przez doktoranta wykazała 17 zmian chromosomalnych, które wystąpiły tylko w zarodkach poronnych.

W celu identyfikacji dodatkowych i utraconych kopii wykorzystano pokrycie odczytami sekwencji genomowej. Pokrycie to jednak było bardzo małe co czyni całkiem sporym prawdopodobieństwo uzyskania przypadkowych wyników. Chciałabym zatem zapytać Doktoranta w jakim stopniu niskie pokrycie mogło wpłynąć na wyniki analizy i czy wykorzystana metoda była w jakiś sposób zoptymalizowana na potrzeby tego projektu?

We wszystkich opisanych powyżej badaniach Bence Galik zaprojektował i wykonał analizy bioinformatyczne. Ze względu na specyficzny typ danych i pytania naukowe każdy projekt wymagał określonej ścieżki analitycznej. Jedynym wyjątkiem są badania pierwsze i drugie, w których można było zastosować to samo podejście. Projektując ścieżki analityczne Doktorant wykorzystał standardowe narzędzia, których wybór jest w pełni uzasadniony. Potrafił także w pełni wykorzystać systemy do zarządzania i obsługi obliczeń. Wykazał tym samym, że zna istniejące narzędzia i potrafi je efektywnie wykorzystać do analiz bioinformatycznych.

Wnioski końcowe

Podsumowując, przedstawiona rozprawa doktorska, jak już wspomniałam, budzi moje wątpliwości ze względów formalnych i etycznych. Jednak ze mając na uwadze jakość przeprowadzonych badań i merytoryczny wkład doktoranta w ważne i ciekawe projekty zdecydowałem się skupić na stronie naukowej. Ocenę kwestii etycznych oraz tego czy przedstawiona praca spełnia warunki określone w Uchwale Senatu UMB 91/2019, załącznik 1f (z 24 października 2019 r.) pozostawiam komisji i Senatowi UMB.

Bence Galik wykonał bardzo solidną pracę, a jego wkład w prezentowane badania był niewątpliwie znaczący. Wykazał również, że posiada wymaganą wiedzę i umiejętności oraz gotowość do podjęcia poważnych zadań badawczych. Niestety nie ustrzegł się błędów redakcyjnych, których w pracy jest całkiem sporo. Ich ilość może sugerować, że praca pisana była w pośpiechu i bez należytego sprawdzenia. Niemniej jednak jakość naukowa prezentowanej rozprawy przeważa nad kwestiami redakcyjnymi i w mojej ocenie rozprawa spełnia warunki określone w art. 187 ustawy z dnia 20 lipca 2018 roku -prawo o szkolnictwie wyższym i nauce (Dziennik Ustaw 2018 poz. 1668) i ustalone w art. 13 ust. 1 ustawy z dnia 14 marca 2003 r. o stopniach i tytule naukowym oraz stopniach i tytule w zakresie sztuki (Dz. U. z 2021 r. poz. 478). W związku z tym rekomenduję dopuszczenie Bence'a Galika do dalszych etapów postępowania.

J. M/

ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ
Faculty of Biology
Institute of Human Biology and Evolution

Izabela Makałowska Ph.D.
Institute of Human Biology and Evolution
Faculty of Biology
Adam Mickiewicz University in Poznań
Poland

Poznań, 3.04.2023

Evaluation of the Doctoral Dissertation prepared by Bence Galik, Msc, a PhD student at the Medical University of Białystok.

Title of the thesis:

"Development and implementation of original bioinformatic pipelines for NGS data analysis in cancer and assisted reproduction research"

In recent years, we have observed a real revolution in life sciences research, this obviously includes biomedical research. This is mostly due to the development of new, high-throughput sequencing technologies, which made genomics and transcriptomics, together with other -omics studies, an integral part of medical research. The massiveness and increasing complexity of projects in which next-generation sequencing is used brings new challenges for bioinformatician. The analysis of results of extensive omics research often forces the simultaneous use of many, not necessarily compatible, bioinformatics tools. The diversity of data types and the specificity of a given studies make it difficult or even impossible, to design universal pipelines for data processing and analysis. Almost every project requires designing a project-specific analytical path. Designing these pipelines and performing required analyses was the main aim of the dissertation prepared by Bence Galik. His research focused on five case studies: chronic lymphocytic leukemia, primary central nervous system lymphoma, glioblastoma multiforme, non-small cell lung cancer, and pre-implantation genetic testing of embryos for in vitro fertilization. The work was carried out under the supervision of dr hab. Attila Gyenesei at the Department of Clinical Molecular Biology of the Medical University of Bialystok.

Formal assessment of the dissertation

Evaluating the presented dissertation in formal terms was a little bit troublesome for me. This was mainly because the anti-plagiarism analysis showed 61% similarity to existing documents. Further analysis clearly indicated that four research cases, out of five constituting the dissertation, were published. Parts describing these studies were just rewritten, with basically insignificant changes, paragraphs of these manuscripts. This accounts for almost 80% of the entire dissertation. The mere use of published results is not a problem. However, the scale is

rather big, and the PhD Candidate didn't mention the fact that these results were published nor that other researchers participated in studies. Thus, in a way, he attributed to himself all the work done. This raises some doubts since all publications were multi-author works, in one case there were as many as 22 co-authors. Therefore, before proceeding with the evaluation, I asked PhD Candidate for a detailed description of his contribution to each of the published studies.

In three of those publications Bence Galik is the equivalent first author, and in my opinion, presenting a collection of these three papers as a dissertation would meet both the statutory requirements and those set out in the resolution of the Senate of the Medical University of Bialystok of October 24, 2019 (resolution 91/2019, Annex 1f). By doing this Bence Galik would avoid abovementioned problems. I would like to ask PhD Candidate to explain this not obvious choice of preparing his thesis.

The dissertation is covering 128 pages of text in English, illustrated with 31 figures and includes 19 tables. The list of cited literature is extensive and contains over 200 items. The layout of the work is typical for doctoral dissertations. The work consists of a summary (in Polish and English), introduction, materials and methods, results, and discussion. The whole is closed with concise conclusions. Unfortunately, the purpose of the work was not clearly specified by the PhD Candidate.

The thesis was written rather correctly, although the PhD Candidate did not avoid some linguistic and editorial imperfections. I will give just a few examples:

- on page 16 there is a reference to [IJC], probably should be a reference to one of the literature items;

- on page 17 it says "Embry" instead of "Embryo"

- on page 20 instead of "in during the run..." should be "during the run..."

- on page 57 it says "Also, other interesting patient specific events were obese during the study". Probably meant "...patient specific events were observed..." not "were obese".

The style of writing is not consistent. Sentences are written sometimes in past, sometimes in present and sometimes in future. An editorial oversight are also illegible descriptions on some figures, for example in Figure 12. Additionally, many abbreviations have not been explained. The work also does not contain a list of abbreviations. In some of described cases there is a lack of information on where the samples come from.

Substantive assessment

In the introduction, Bence Galik focused primarily on discussing individual case studies and related diseases. Unfortunately, there is not much about bioinformatics, high-throughput sequencing technologies and their applications, as well as existing tools. These topics are discussed very briefly and, in my opinion, in a way that is not entirely satisfactory, given the bioinformatic nature of the dissertation.

The first study case refers to chronic lymphocytic leukemia (CLL). The goal of performed analyses was to dissect the clonal evolution that affects 30 genes recurrently mutated in CLL. The study was conducted in the context of ibrutinib therapy based on targeted sequencing of the genes of interest. Peripheral blood samples from 20 patients were collected before and after treatment. As a result of the analyses, the Bence Galik identified a total of 211 somatic variants and showed that the number of mutations in the samples increased after treatment. He also found that as many as 40% of genes have multiple mutations. Interestingly, clones carrying mutations in the *IGLL5, EIF2A* and *EP300* genes were eliminated by treatment. In

addition, the PhD student described a new pattern of clonal mutation dynamics in the *BTK* and *TP53* genes. Obtained results are very interesting as they demonstrate that comprehensive deep sequencing of driver genes and analysis of clonal evolution provide clinical benefits. This is quite important for achieving better characterization of CLL in individual patients and in the consequence better, more personalized treatment.

The second research case involved primary central nervous system lymphoma (PCNSL). As with CLL, targeted amplicon sequencing was performed, with the difference that DNA was extracted from FFPE samples. The material came from 64 patients with primary lymphoma and 12 patients with secondary lymphoma (SCNSL). Mutation profiles were determined by the PhD Candidate for 14 genes. The analysis carried out revealed 239 mutations in all lymphomas combined. Of these, 210 were seen in patients with primary lymphoma. The genes with the most mutations in PCSNL were *MYD88, PIM1, KMT2D* and *PRDM1*. However, no mutations in the *PTPRD* gene were observed. In the secondary lymphoma cohort, mutations were most common in the *PRDM1, MYD88* and *PIM1* genes. No mutations have been identified in the *CARD11, CSMD2, CSMD3* and *PTPRD* genes. Bence Galik then compared the differences in mutation distribution in patients with primary and secondary lymphoma, as well as between the ABC and GC subtypes. However, no statistically significant differences were found.

I have two questions relevant to this study case:

1. In the description of bioinformatic workflow there is a reference to individual sample-specific libraries A and B that were combined (page 35). However, there is no explanation on what A and B libraries are. Could this be explained?
2. Patients' sub-classification is compared based on two methods. However, only LST-assay is described in the methods section. Hans's algorithm was for the first time mentioned in the description of results. Figure demonstrating the algorithm is not informative, especially for non-experts in lymphoma classification. I would like to ask for better explanation on how this algorithm works.

In subsequent studies, the PhD Candidate analyzed data on glioblastoma multiforme (GBM), which is a very aggressive brain tumor. The aim of the study was to identify the molecular factors associated with the development and recurrence of GBM. Here, the analysis was focused on DNA methylation patterns. DNA CpG methylation patterns were determined based on the sequencing of bisulfite converted libraries. Bence Galik compared methylation patterns in controls, tumors before chemoradiation treatment, and recurrent tumors after therapy. DNA was derived from FFPE samples of 22 patients. As a control data deposited in the EGA database (The European Genome-phenome Archive) was used. These were samples taken during the epilepsy surgery. Comparison of the methylation level demonstrated a shift toward hypomethylation - the highest methylation level in the control, and the lowest in recurrent tumors. However, comparisons of differential methylation data at site and region levels revealed no significant *p* values in any of the three pairwise comparisons. On the other hand, GO analyzes allowed the identification of several pathways with biological relevance. Based on the obtained data, a mechanism that may underlie the development and progression of glioblastoma has been proposed. The obtained results, and in particular the proposed mechanism, may be important for the development of new treatment methods.

In this study samples from two different sources were used. This is generating technical differences, which could and most probably did affect results of the analysis. Has the impact of this so-called batch effect assessed, and any techniques used to minimalize this impact?

The fourth project was focused on miRNA expression in non-small cell lung cancer (NSCLC). Tumor, peripheral blood, and control samples from patients diagnosed with lung adenocarcinoma (AC) and squamous cell carcinoma (SCC) were analyzed. The PhD Candidate

managed to identify over 1,600 miRNAs across all samples. To compare individual samples and identify outliers, the correlation between samples was calculated. Sample similarity was also checked by hierarchical clustering and principal component analysis. In the dendrogram, the control samples clearly separated from the tumor samples. Principal component analysis showed a clear separation between adenocarcinoma samples and its control as well as between squamous cell carcinoma and control. Interestingly, the adenocarcinoma samples overlapped with the squamous cell carcinoma control and the squamous cell carcinoma samples with the adenocarcinoma control. This result is quite intriguing. Could it be explained somehow?

The subsequent differential expression analysis allowed the PhD Candidate to identify miRNAs with over- and under-expression in both types of cancer, as well as miRNAs with type-specific changes. Further KEEG analysis revealed microRNA-related metabolic pathways underlying the pathogenesis of lung adenocarcinoma and squamous cell carcinoma. The last step was to build a model based on the 17 miRNAs that showed the greatest changes in expression and are potential biomarkers. The model was then validated on a set of samples from peripheral blood. The validation results were not entirely satisfactory, which is probably because the validation set was not the best for these studies. This was noticed also by PhD candidate. Therefore, I wanted to ask why this particular dataset was selected for model validation?

The aim of the last project was to develop a bioinformatics pipeline for non-invasive preimplantation genetic testing (NIPGT-A). A total of 28 samples were sequenced and 22 were selected for further analysis based on sequencing quality. The analysis included five samples of control medium, two samples of cord blood with known CNVs and 15 samples of culture medium of 3-day-old embryos. Six samples were from miscarriages and nine were from successful pregnancies. Following genomic DNA sequencing, copy number variants were identified. The analysis carried out by the PhD Candidate showed 17 chromosomal changes that occurred only in abortive embryos. Genomic sequence read coverage was used to identify copy gains and losses. However, this coverage was very low, which increases the likelihood of random results. Therefore, I would like to ask the PhD Candidate to what extent the low coverage could have affected the results of the analysis and whether the method used was somehow optimized for the needs of this project?

In all the studies described above, Bence Galik designed and performed bioinformatics analyses. Due to the specific type of data and scientific questions, each project required a specific analytical pipeline. The only exceptions are studies one and two, where the same approach could be used. Tools that were incorporated into pipelines were standard and their selection fully justified. The PhD candidate took advantage of workflow management system as well as Docker and Singularity technology for the multi-scale handling of containerized computation. This demonstrated that Bence Galik is familiar with existing tools and knows how to utilize them to build efficient pipelines for bioinformatic analyses.

Final remarks

To sum up, the presented doctoral dissertation, as I have already mentioned, raises some doubts for formal and ethical reasons. However, due to the quality of the research and the substantive contribution of the PhD Candidate to important and interesting projects, I decided to focus on the scientific side. I leave the assessment of ethical issues and whether the submitted work meets the conditions set out in the Resolution of the Senate of the MUB 91/2019, Annex 1f (of October 24, 2019) to the committee and the Senate of the MUB.

Bence Galik has done a very solid job and his contribution to the presented research was undoubtedly significant. He has also demonstrated that he has the required knowledge and skills and is ready to undertake serious research tasks. Unfortunately, he did not avoid editorial errors, of which there are quite a lot in the work. Their number may suggest that the thesis was written in a hurry and without proper checking. Nevertheless, the scientific quality of the presented dissertation outweighs editorial issues and in my opinion the dissertation meets the conditions set out in Art. 187 of the Act of July 20, 2018 - law on higher education and science (Journal of Laws 2018, item 1668) and set out in art. 13 sec. 1 of the Act of 14 March 2003 on academic degrees and titles and degrees and titles in the field of art (Journal of Laws of 2021, item 478). Therefore, I recommend admitting Bence Galik to further stages of the proceedings.

J. M/